UNICAMP

UNIVERSIDADE ESTADUAL DE CAMPINAS

Instituto de Matemática, Estatística e
Computação Científica

ANDRÉ FELIPE BERDUSCO MENEZES

Métodos estatísticos para modelagem de dados de scRNA-seq

Campinas

2021

André Felipe Berdusco Menezes

Métodos estatísticos para modelagem de dados de scRNA-seq

Dissertação apresentada ao Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestre em Estatística.

Orientador: Benilton de Sá Carvalho

Este exemplar corresponde à versão final da Dissertação defendida pelo aluno André Felipe Berdusco Menezes e orientada pelo Prof. Dr. Benilton de Sá Carvalho.

Campinas

2021

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

M524m

Menezes, André Felipe Berdusco, 1996-
Métodos estatísticos para modelagem de dados de scRNA-seq / André Felipe Berdusco Menezes. – Campinas, SP : [s.n.], 2021.

Orientador: Benilton de Sá Carvalho.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica.

1. scRNA-seq. 2. Dados omics. 3. Expressão diferencial. 4. Análise de agrupamento. 5. Bioestatística. I. Carvalho, Benilton de Sá, 1979-. II. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Statistical methods for scRNA-seq data modeling
Palavras-chave em inglês:
scRNA-seq
Omics data
Differential expression
Cluster analysis
Biostatistics
Área de concentração: Estatística
Titulação: Mestre em Estatística
Banca examinadora:
Benilton de Sá Carvalho [Orientador]
Samara Flamini Kiihl
Diogo Fernando Troggian Veiga
Data de defesa: 15-04-2021
Programa de Pós-Graduação: Estatística

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0002-3320-9834
- Currículo Lattes do autor: http://lattes.cnpq.br/3911619582088894

Dissertação de Mestrado defendida em 15 de abril de 2021 e aprovada pela banca examinadora composta pelos Profs. Drs.

Prof(a). Dr(a). BENILTON DE SÁ CARVALHO

Prof(a). Dr(a). SAMARA FLAMINI KIIHL

Prof(a). Dr(a). DIOGO FERNANDO TROGGIAN VEIGA

A Ata da Defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria de Pós-Graduação do Instituto de Matemática, Estatística e Computação Científica.

"Aos meus queridos pais"

Agradecimentos

Aos meus pais que me permitiram um ensino de qualidade, pelo apoio incondicional, pelo amor e carinho que sempre encontrei em vocês.

Ao meu orientador, prof. Dr. Benilton de Sá Carvalho, pelos ensinamentos, orientações e por sua paciência nos meus momentos de nervosismo.

Ao Josmar Mazucheli, pela parceria e contribuição inestimável na minha formação acadêmica durante minha graduação.

Aos meus amigos, Victor, Wesley e Gabriel, pelo companheirismo e trocas de experiências que contribuíram para minha formação pessoal e acadêmica.

Aos colegas que pude encontrar durante o mestrado. De forma especial, ao Andreson, parceiro das cervejas e dos sambas. Ao Rafael, pelas conversas e reflexões. Ao Henrique, pelas explicações extremamente didáticas sobre imunologia.

Aos funcionários do IMECC, especialmente a dona Zefa pelos cafezinhos e conversas, meu muito obrigado. Ao amigo Reginaldo pelas conversas do dia a dia.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico — CNPq (132278/2019-7), pelo apoio financeiro.

Com que medidas se medem
Esses seres dos acasos
Que se cruzam por aí
Predestinados
A tudo ou nada?
Por acaso sou eu um?
(Jandyra Waters, Pedras Nuas)

Resumo

A tecnologia de sequenciamento massivo de DNA e RNA permitiu desenvolvimentos significativos na área de biomédicas, culminando, neste momento, na implementação da medicina de precisão, em que informações moleculares do paciente são levadas em consideração para a tomada de decisão no curso de tratamento. Em geral, o sequenciamento de DNA ou RNA é realizado a partir de uma amostra do tipo *bulk*, em que o material molecular de várias células é combinado para a realização das análises de interesse. Com o avanço tecnológico mais recente, tornou-se possível o sequenciamento do material genético no nível das células, permitindo a análise de fenótipos de interesse, como doenças diversas, numa resolução ainda mais granular que aquela oferecida por amostras do tipo *bulk*. Especificamente, a tecnologia de *single-cell* RNA sequencing (scRNA-seq) permite criar perfis de expressão na resolução da célula. A vasta quantidade de dados produzida por experimentos de scRNA-seq e as hipóteses de pesquisa que os motivam exigem um tratamento computacional e estatístico eficiente. Nesse sentido, o objetivo dessa pesquisa foi estudar as técnicas utilizadas para análise dos dados em experimentos scRNA-seq, que incluem (i) métodos para pré-processamento, (ii) processamento da matriz de contagem e (iii) métodos estatísticos para análise dos dados. Além disso, motivado por um conjunto de dados de células do tecido BALF de pacientes com COVID-19, um estudo de simulação foi conduzido considerando as características particulares dos dados para comparar diferentes abordagens para análise de expressão diferencial que incorporam a origem da célula. Por fim, o fluxo usual de análise discutido no trabalho foi empregado para analisar o conjunto de dados de células BALF, caracterizando grupos de células e comparando os níveis de expressão gênica dos indivíduos sob diferentes condições experimentais.

Palavras-chave: scRNA-seq, dados *omics*, expressão diferencial, análise de agrupamento, bioestatística.

Abstract

High-throughput sequencing technology allowed significant developments in the biomedical area, culminating, at this moment, in the implementation of precision medicine, where the patient's molecular information is taken into consideration for decision making in the treatment course. In general, the DNA or RNA sequencing is performed from bulk sample, where the molecular material of several cells is combined to perform the analysis of interested. Especially, the technology of single cell RNA sequencing (scRNA-seq) enables high-throughput transcriptome profiling at the resolution of single cells. The vast amount of data produced by scRNA-seq experiments and the research hypotheses that motivate them require efficient computational and statistical treatment. Hence, the goal of this research was to study the techniques used for data analysis in scRNA-seq, which include (i) methods for pre-processing raw data, (ii) data processing of counting matrix and (iii) statistical methods for data analysis. In addition, motivated by a data set of cells from bronchoalveolar lavage fluid (BALF) tissue from patients with COVID-19, a simulation study was conducted, considering the particularities of the data, to compare different approaches for differential expression analysis that incorporate the cell's origin. Finally, the usual workflow discussed in the research was adopted to analyze the BALF cells data set by characterizing groups of cells and comparing the expression genes levels of individuals under different experimental conditions.

Keywords: scRNA-seq, omics data, differential expression, cluster analysis, biostatistics.

Lista de ilustrações

Lista de tabelas

Sumário

Introdução

O dogma central da biologia molecular introduzido por Crick (1970) descreve o fluxo da informação genética. Resumidamente, a informação contida no DNA é transcrita para o RNA e codificada em proteínas. As proteínas, por sua vez, conferem ao indivíduo o fenótipo observável, como, por exemplo, uma doença. Uma estratégia custo-efetiva para o entendimento das funções proteicas é estudar a expressão gênica nos cenários de interesse, por exemplo, em um grupo de indivíduos saudáveis, comparando os resultados com aqueles obtidos em um grupo de indivíduos acometidos de uma doença. A análise simultânea dos aproximadamente 20.000 genes humanos combinada a metodologias de análise de dados tem possibilitado a identificação de assinaturas moleculares preditoras do fenótipo em questão (KUKURBA; MONTGOMERY, 2015).

A análise simultânea da expressão gênica para milhares de alvos moleculares (genes) só é possível pelo avanço nas técnicas de sequenciamento massivamente paralelo (*high-throughput sequencing*) (REUTER; SPACEK; SNYDER, 2015). Esta estratégia permite que pesquisadores coletem amostras de um tecido de interesse em vários organismos, habitualmente distribuídos em, pelo menos, dois ou mais grupos, por exemplo, saudáveis versus afetados. Para cada uma das amostras, extrai-se o RNA das inúmeras células ali presentes. O RNA coletado deste conjunto de células é sequenciado para cada organismo amostrado e alinhado a algum genoma de referência, permitindo assim a caracterização da expressão gênica. Este tipo de sequenciamento, conhecido como *bulk* RNA *sequencing* (RNA-seq), surgiu na década de 2000, com os primeiros artigos científicos publicados por Cheung et al. (2006), Weber et al. (2007) e Emrich et al. (2007). Embora a tecnologia de RNA-seq tenha possibilitado importantes descobertas científicas de diferentes eventos biológicos analisados (WANG; GERSTEIN; SNYDER, 2009; SERRATÌ et al., 2016), o RNA extraído e estudado representa uma média de milhares a milhões de transcriptomas de células presentes na amostra, assim, diferenças potencialmente significativas e biologicamente importantes entre as células podem ser obscurecidas (OLSEN; BARYAWNO, 2018).

Nesse contexto, Tang et al. (2009) apresentaram um avanço incrível na área de quantificação da expressão gênica, a equipe demonstrou como isolar e sequenciar o RNA de uma única célula. Nos primeiros estudos de *single-cell* RNA *sequencing* (scRNA-seq) de 10 a 100 células foram analisadas e caracterizadas (PICELLI et al., 2013; SHALEK et al., 2013; KUMAR et al., 2014). Desde então, a técnica tem sido gradualmente melhorada a custos acessíveis com possibilidades de analisar na ordem de milhões de células, por exemplo, Zheng et al. (2017) analisaram cerca de 1,3 milhões de células cerebrais. A importância científica do scRNA-seq foi reconhecida pela revista Nature em 2013 como o "Método do Ano" (METHODS, 2013).

Experimentos de scRNA-seq permitem que medições da expressão gênica sejam realizadas numa resolução com maior granularidade antes não obtida, a célula, possibilitando que grupos de tipo de células sejam distinguidos, o arranjo das populações de células de acordo com para novas hierarquias e a identificação de células em transição entre estados. Este tipo de tecnologia permite uma visão muito mais clara da dinâmica do desenvolvimento do tecido ou organismo e das estruturas dentro das populações de células que até então eram assumidas como homogêneas, especialmente em experimentos de *bulk* RNA-seq (LAHNEMANN et al., 2020).

Os dados gerados por experimentos de scRNA-seq fornecem oportunidades importantes para o avanço da ciência. Conforme discutem Lahnemann et al. (2020), só agora é possível reavaliar as hipóteses sobre as diferenças entre os grupos de amostra predefinidos no nível da célula, não importando se esses grupos de amostras são subtipos de doenças, grupos de tratamento ou simplesmente tipos de células morfologicamente distintos. Um exemplo interessante é o *Human Cell Atlas* (REGEV et al., 2017), uma iniciativa que visa mapear os inúmeros tipos de células e estados que compõem um ser humano.

Atreladas às oportunidades únicas fornecidas por experimentos de scRNA-seq, estão as características particulares antes não observáveis em outros tipos de sequenciamento. Quantidades limitadas de material disponível por célula levam a altos níveis de incerteza sobre as observações. Quando a amplificação é usada para gerar mais material, ruído técnico é adicionado aos dados resultantes. Dessa forma, análises para controle de qualidade das células amostradas e métodos específicos para normalização das expressões gênicas são etapas importantes na análise dos dados de scRNA-seq (LUN; BACH; MARIONI, 2016; VALLEJOS et al., 2017).

A análise de expressão diferencial (DE) tem como objetivo identificar e quantificar genes que apresentem níveis de expressão diferentes entre amostras sob certas condições experimentais. No contexto de scRNA-seq, os métodos desenvolvidos para análise DE assumem que os grupos de células a serem comparados sejam homogêneos, todavia, devido a grande variabilidade entre as células, nos fluxos usuais de análise é realizado previamente o agrupamento das células (KISELEV; ANDREWS; HEMBERG, 2019) ou anotação dos tipos das células (ARAN et al., 2019) para identificar subpopulações de células e mitigar a heterogeneidade presente. Assim, a análise DE é conduzida incorporando a variabilidade intrínseca das subpopulações de células identificas, por exemplo, comparando as células de diferentes condições experimentais dentro de cada subpopulação (AMEZQUITA et al., 2020).

À medida em que as tecnologias de scRNA-seq tem se desenvolvido, conjuntos de dados complexos tem se tornado disponíveis em vários repositórios públicos (SONESON; ROBINSON, 2018; AMEZQUITA et al., 2020). Essa vasta quantidade de dados e as

hipóteses de pesquisas que os motivam exigem um tratamento computacionalmente eficiente e estatisticamente rigoroso (AMEZQUITA et al., 2020; LAHNEMANN et al., 2020). Esses aspectos correspondem a uma definição de "Ciência de Dados" discutido por Hicks e Peng (2019) e postulada por Lahnemann et al. (2020) como a era "*single-cell data science*".

A importância contemporânea dos experimentos de scRNA-seq tem sido acompanhada pela comunidade científica da bioinformática, que vem desenvolvendo e disponibilizando diversos métodos para processamento e análise dos dados de scRNA-seq. Zappia, Phipson e Oshlack (2018) desenvolveram o banco de dados scRNA-tools (<www.scRNA-tools.org>) que cataloga as ferramentas computacionais disponíveis para análise de scRNA-seq. Os autores apresentam um catálogo dos métodos conforme o *software* e categoria de análise na qual a ferramenta se enquadra. A evolução temporal dos métodos disponíveis e dos softwares com maiores implementações catalogados pelo scRNA-tools é ilustrado na Figura 0.1. As linguagens R e Python tem sido os softwares com maior número de métodos implementados. O projeto Bioconductor tem contribuído ativamente a comunidade científica para difusão dos métodos de scRNA-seq no ambiente R, fornecendo uma experiência atraente aos usuários com documentações detalhadas e código aberto (HUBER et al., 2015), o Bioconductor tem se destacado com principal fonte dos recentes métodos propostos para scRNA-seq (AMEZQUITA et al., 2020).

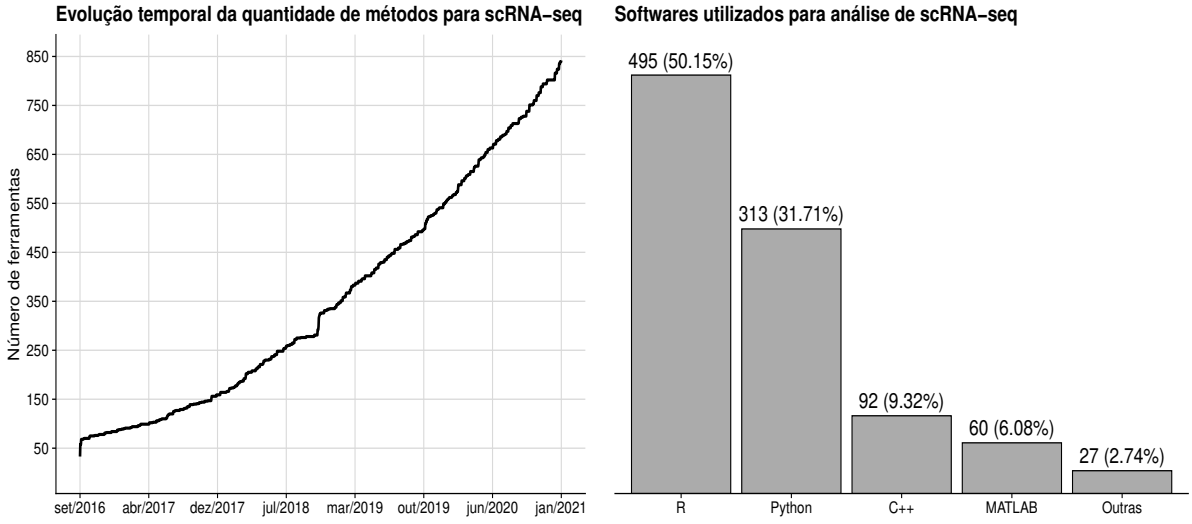


Figura 0.1 – Informações obtidas pelo banco de dados scRNA-tools sobre a evolução temporal dos métodos e *softwares* utilizados para análise de scRNA-seq.

Tendo em vista o panorama discutido, o presente trabalho tem como objetivo estudar as principais técnicas estatísticas para análise de dados provenientes de experimentos scRNA-seq. O referencial teórico, portanto, foram livros e artigos científicos dos notáveis pesquisadores da área, como Crowell et al. (2020), Amezquita et al. (2020), Soneson e Robinson (2018), Lun e Marioni (2017), Lun, Bach e Marioni (2016), McCarthy,

Chen e Smyth (2012), que discutem e apresentam conceitos metodológicos sobre a analise de dados de scRNA-seq.

Para melhor sistematização dos estudos realizados, a pesquisa está organizada em quatro capítulos que se complementam. O Capítulo 1 aborda uma revisão sobre conceitos da biologia molecular da célula, discussões sobre o sequenciamento genético, as etapas experimentais e o pré-processamento dos dados de scRNA-seq.

O Capítulo 2 abarca sobre as etapas de processamento da matriz de contagem, especificamente, definições e discussões metodológicas sobre abordagens para avaliar o controle de qualidade das contagens, métodos de normalização e transformação logarítmica das expressões gênicas, métodos para seleção inicial das *features* e técnicas de redução de dimensionalidade.

Os métodos estatísticos utilizados na análise dos dados de scRNA-seq são discutidos no Capítulo 3. Particularmente, este capítulo traz definições sobre o método de agrupamento k-means juntamente com técnicas para seleção do número de grupos e avaliação da estabilidade dos grupos. Os principais testes estatísticos de comparação múltipla para detecção de genes marcadores sugeridos por Lun, McCarthy e Marioni (2016) são revisados e discutidos. Os principais métodos e pacotes do Bioconductor para realizar a anotação dos tipos de células são apresentados. Além disso, uma revisão dos modelos estatísticos para análise da expressão gênica e discussões das definições teóricas dos modelos edgeR (ROBINSON; MCCARTHY; SMYTH, 2009) e linear misto no contexto de análise diferencial são também abordados.

Como forma de aplicar os conceitos discutidos ao longo da dissertação o Capítulo 4 apresenta um estudo de simulação e uma análise de dados reais. Em particular, o estudo de simulação é conduzido para comparar propriedades estatísticas dos modelos discutidos na análise de expressão gênica quando as células sequenciadas são oriundas de diferentes organismos, por exemplo, indivíduos. A análise de dados reais é conduzida utilizando o conjunto de dados de Liao et al. (2020) e os objetivos foram caracterizar as células e comparar o nível de expressão gênica entre as condições dos indivíduos.

1 Célula e o Sequenciamento Genético

Para melhor caracterização do objeto de pesquisa, este capítulo apresenta na Seção 1.1 uma discussão dos principais conceitos da biologia molecular da célula. Em seguida, definições e discussão do sequenciamento genético do RNA são abordados na Seção 1.2. Finalmente, as Seções 1.2.1 e 1.2.2 discutem, respectivamente, as características das etapas experimentais do scRNA-seq e o pré-processamento dos dados.

1.1 Biologia Molecular da Célula

Mesmo com a aparente diversidade, os seres vivos possuem grandes similaridades. Todo organismo vivo é composto por uma ou mais células. Muitas formas de vida consistem em uma única célula, por exemplo o fermento. Outros seres vivos são compostos por vastos grupos celulares que desempenham distintas funções, por exemplo o ser humano. A célula é uma estrutura microscópica considerada a unidade fundamental da vida. Células possuem a capacidade de gerar energia, crescer e se dividir transferindo suas características entre si, assim, novas células resultam da divisão de alguma célula pré-existente.

A estrutura da célula é formada por uma membrana que separa o interior da célula, o citoplasma, do ambiente exterior. A membrana é composta por fosfolipídios e proteínas. Muitas das reações envolvidas na quebra de nutrientes para liberar energia são catalizadas por enzimas localizadas no citoplasma. Em particular, estão os ribossomos que são organelas responsáveis pele a síntese proteica, isto é, a produção de proteína (CLARK; PAZDERNIK; MCGEHEE, 2019). Uma classificação usual dos organismos vivos ocorre conforme as diferenças de compartimentação da célula. Aqueles cujas células possuem núcleo são denominados eucariontes, enquanto que os procarionte são aqueles que as células não possuem núcleo (ver Figura 1.1).

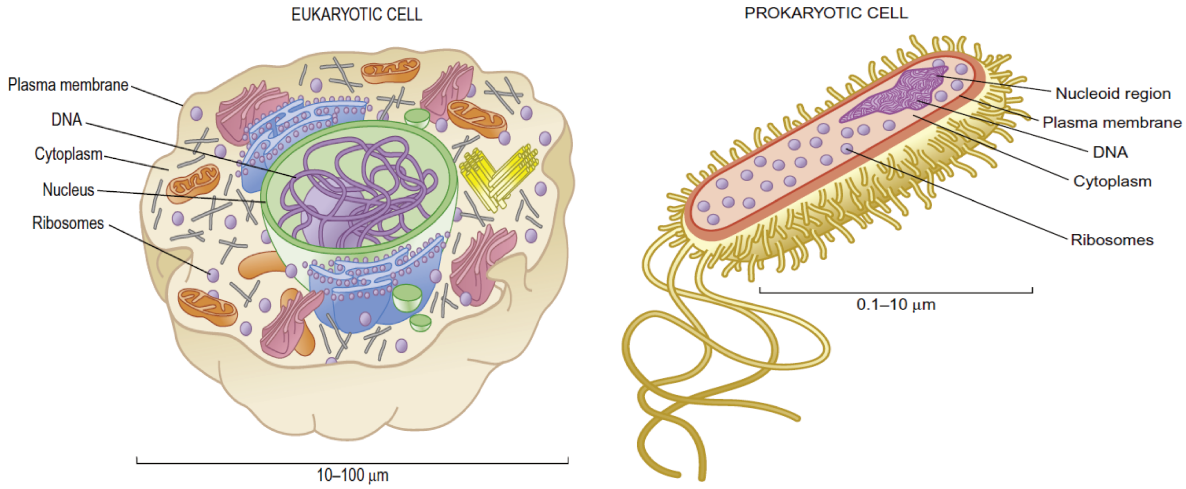


Figura 1.1 – Estrutura das células de organismos eucarionte e procarionte. Imagem retirada de Clark, Pazdernik e McGehee (2019).

Todo organismo vivo armazena na célula sua informação hereditária no formato de moléculas de ácido desoxirribonucleico (*deoxyribonucleic acid* – DNA). O DNA é uma molécula de estrutura helicoidal (dupla-hélice) formada por cadeias longas de polímeros emparelhadas e compostas sempre com os mesmos quatro tipos de monômeros, denominados nucleotídeos (ALBERTS et al., 2015). Em uma única fita do DNA cada nucleotídeo consiste em um açúcar (desoxirribose) com um grupo fosfato ligado a ele e uma base nitrogenada que pode ser adenina (A), guanina (G), citosina (C) ou timina (T). A união entre as duas cadeias ocorre pelo pareamento das bases seguindo a regra: A liga com T e C liga com G.

Para desempenhar a função de transmissão genética, o DNA precisa sintetizar outras moléculas na célula. A transferência de informação genética inicia-se sintetizando segmentos da sequência de DNA em moléculas de ácido ribonucleico (*ribonucleic acid* – RNA), este processo é denominado transcrição. Na sequência, em um processo mais complexo denominado tradução, os ribossomos decodificam a informação genética contida no RNA para a síntese proteica (ALBERTS et al., 2015). Resumidamente, a informação genética é armazenada em segmentos de DNA, transcritas em RNA e codificadas em proteínas. Este princípio, denominado de dogma central da biologia molecular foi proposto por Crick (1970) e ocorre em todos seres vivos, de bactérias a seres humanos.

A molécula de RNA é similar ao DNA com a diferença que o RNA é uma fita única, o açúcar desoxirribose é substituído pela ribose e a base de uracila (U) é utilizada no lugar de timina (T). É comum utilizar a expressão poly(A)-tail (cauda de ácido poliadenítico) para abreviar os nucleotídeos do RNA, com a letra para a base que o nucleotídeo contém. Por convenção, as sequências de RNA são escritas na direção de 5' a 3'. A Figura 1.2 ilustra a estrutura química do DNA e RNA.

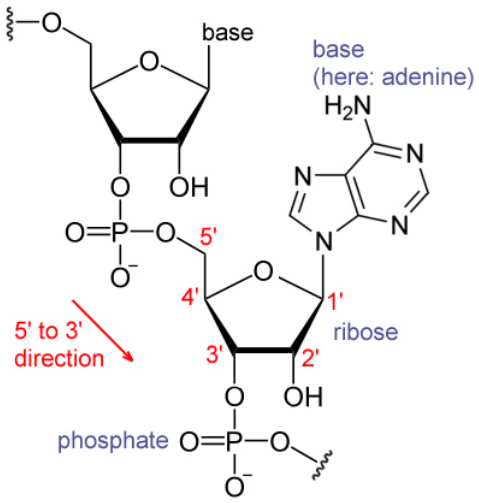


Figura 1.2 – Estrutura química do RNA. Imagem retirada de <https://commons.wikimedia.org/wiki/File:RNA_chemical_structure_adenine.JPG>.

Existem diferentes tipos de RNA que realizam distintas funções na célula. Moléculas de RNA sintetizadas em proteínas são os RNA mensageiros (mRNA). Outras principais classes de RNA incluem o RNA de transferência (tRNA) responsável pelo transporte das moléculas de aminoácidos até os ribossomos e o RNA ribossômico (rRNA) que compõe parte do ribossomo e auxilia na montagem da cadeia de aminoácidos pelo mRNA e tRNA.

As proteínas, por sua vez, são constituídas por uma ou mais cadeias de aminoácidos, sendo que existem 20 tipos de aminoácidos. As moléculas de proteínas são indispensáveis para o funcionamento da célula e desempenham diferentes atividades conforme a sequência de aminoácidos que as compõem.

Segmentos de DNA são agrupados em unidades conhecidas como genes. Conforme Alberts et al. (2015), o gene é uma sequência de segmentos do DNA correspondente a uma única proteína ou conjunto de proteínas ou, ainda, como sendo uma molécula de RNA catalítico, regulatório ou estrutural.

A regulação da expressão gênica ocorre durante a transcrição e processamento de RNA que varia conforme o tecido. Por exemplo, no que se refere à regulação gênica, a transcrição na pele difere da transcrição no cérebro. Por esta natureza, a expressão gênica é habitualmente utilizada para identificação de padrões que diferenciam, não apenas, tecidos diferentes, mas também para comparar o mesmo tecido entre indivíduos sob diferentes condições. Por exemplo, pacientes com câncer de estômago versus indivíduos saudáveis.

1.2 Sequenciamento do RNA

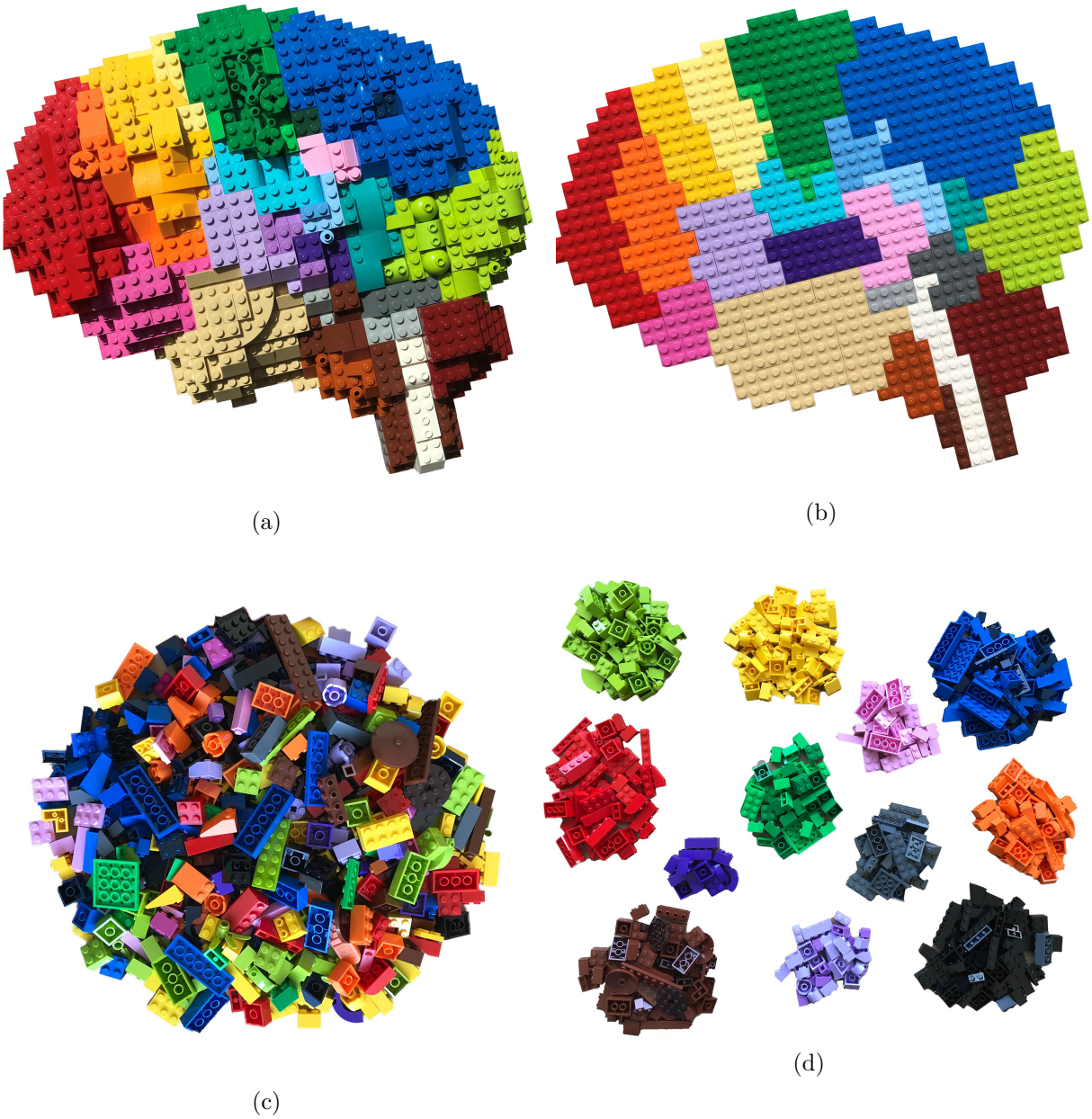
Com o advento tecnológico das técnicas de sequenciamento massivamente paralelo (*high-throughput sequencing*), tem sido possível analisar simultaneamente a expressão gênica para milhares de alvos moleculares (genes). Esta estratégia permite que pesquisadores ou médicos colham amostras de um tecido de interesse em vários indivíduos, habitualmente distribuídos em, pelo menos, dois grupos (como saudáveis versus afetados, por exemplo). Para cada uma das amostras, extrai-se o RNA das inúmeras células ali presentes. O RNA colhido deste conjunto de células é sequenciado para cada indivíduo amostrado, permitindo a caracterização do perfil de expressão gênica. A análise de expressão gênica, num espaço altamente dimensional, tem permitido um melhor entendimento dos eventos biológicos analisados e provido meios de implementar a medicina de precisão no serviço médico de vários países. Este tipo de sequenciamento é denominado *bulk* RNA-seq (WANG; GERSTEIN; SNYDER, 2009).

Tang et al. (2009) apresentam um avanço incrível na área de quantificação da expressão gênica. A equipe demonstrou como isolar e sequenciar o RNA de uma única célula e não de um *pool* de células, como vinha acontecendo até então. Esta técnica, *single-cell* RNA *sequencing* ou, simplesmente, scRNA-seq, aumentou excepcionalmente a granularidade em que as análises podem acontecer. Possibilitando, assim, a análise das diferenças nas expressões gênicas entre as células e a inferência de suas funções dentro do microambiente a que pertence. Isso pode permitir, por exemplo, a identificação de grupos celulares raros com funções extremamente especializadas no desenvolvimento de uma doença.

A análise do transcriptoma no nível da célula através de experimentos de scRNA-seq permitiu responder muitas questões científicas que não podiam ser respondidas com experimentos de *bulk* RNA-seq. Por exemplo, os tumores humanos são compostos por diversas células e sua caracterização precisa não é alcançada através de sequenciamentos do tipo *bulk* RNA (SUVÀ; TIROSH, 2019). Entretanto, estudos utilizando scRNA-seq tem permitido identificar padrões em mutações somáticas causadas por diferentes processos mutacionais que levam o desenvolvimento de tumores (ALEXANDROV; STRATTON, 2014). A revisão de Tirosh e Suvà (2019) aponta diversos estudos inicias sobre tumores humanos conduzidos com o sequenciamento em célula individual. Além disso, devido à caracterização granular que esta tecnologia oferece, tem sido possível isolar, caracterizar e até descobrir novos microrganismos complexos presentes no meio ambiente (BLAINEY, 2013).

Para ilustrar didaticamente as diferenças entre os sequenciamentos *bulk* RNA-seq e scRNA-seq, pesquisador Bo Xia elaborou uma representação utilizando lego dos sequenciamentos. A Figura 1.3a representa o órgão original de interesse, onde cada pedaço do lego refere-se a uma célula e está colorida conforme o tipo. A Figura 1.3b ilustra

uma visualização espacial do transcriptoma de interesse. No *bulk* RNA-seq as células são misturadas e tratadas como sendo de uma única população e o resultado do sequenciamento é sumarizado entre o conjunto de célula, ou seja, não é possível distinguir as subpopulações presentes (ver Figura 1.3c). Em contrapartida, no scRNA-seq cada célula é sequenciada de forma individual, possibilitando que as diferenças entre os tipos de células sejam observáveis (ver Figura 1.3d).



(a) (b)

(c) (d)

Figura 1.3 – Representação em lego comparando os sequenciamentos *bulk* RNA-seq e scRNA-seq nas células de um determinado tecido. Imagens construídas por Bo Xia e disponíveis em <https://twitter.com/BoXia7/status/1261464021322137600>. (1.3a) Órgão original. (1.3b) Visualização espacial do transcriptoma. (1.3c) Sequenciamento *bulk* RNA-seq. (1.3d) Sequenciamento scRNA-seq.

1.2.1 Etapas Experimentais do scRNA-seq

Estudos de scRNA-seq apresentam várias etapas experimentais que devem ser cuidadosamente empregadas para obtenção dos dados. De forma geral, os procedimentos experimentais consistem em: (i) dissociação da célula; (ii) isolamento da célula; (iii) captura

da molécula de RNA; (iv) transcrição reversa e amplificação; (v) preparação da biblioteca (*library*) e (vi) sequenciamento.

O material de entrada para experimentos scRNA-seq são, em geral, amostras de tecidos biológicos. Inicialmente, este tecido deve ser dissociado em suas células constituintes. O processo de dissociação da célula ocorre por meio de processos mecânicos e enzimáticos. Conforme Lafzi et al. (2018) descrevem, o tecido é dissociado por corte mecânico ou via lâminas. Em seguida, enzimas de digestão são utilizadas para separar as células. Lafzi et al. (2018) fornecem recomendações gerais de enzimas, tempo e temperatura de digestão para diferentes tecidos em humanos e ratos, incluindo, por exemplo, o fígado, pulmão, pâncreas, entre outros.

Para o isolamento da célula não há método padrão e vários esforços tem sido realizados para desenvolvimento e refinamentos dos métodos. No nível mais rudimentar, a célula pode ser isolada manualmente via micropipetagem ou microdissecção (SALIBA et al., 2014). Mesmo apresentando baixo custo-benefício essas técnicas podem ser utilizadas em situações de amostras com poucas células ou amostras raras.

Alternativas de alto rendimento permitem que centenas de células sejam processadas simultaneamente minimizando esforços e custos. Nesta classe destacam-se as plataformas que utilizam FACS (*fluorescence-activated cell sorting*) ou microfluidos (SHAPIRO; BIEZUNER; LINNARSSON, 2013). A primeira plataforma comercial disponível, Fluidigm C1, permitiu o sequenciamento de até 96 células por chip, porém dispositivos mais recentes são capazes de capturar até 800 células (FLUIDIGM, 2020). Uma limitação dessa abordagem, conhecida como *microfluidic plate-based*, é que funciona para células de tamanho relativamente homogêneo (KOLODZIEJCZYK et al., 2015).

Outro método de alto rendimento bastante utilizado é conhecido como *microfluidic droplet-based*. Essa abordagem possibilita a captura de milhares de células, sendo também menos seletivo em relação ao tamanho das células. No entanto, a principal vantagem deste método é a presença de sequências aleatórias de nucleotídeos conhecidos como *Unique Molecular Identifiers* (UMI) no momento do sequenciamento, permitindo reduzir variações técnicas do experimento (LAFZI et al., 2018).

A popularização dos métodos *droplet-based* ocorreu com as plataformas Drop-seq (MACOSKO et al., 2015) e InDrop (KLEIN et al., 2015). A plataforma comercial 10X Genomics Chromium (ZHENG et al., 2017) também tem ganhado relevância nesta classe de métodos. O recente estudo conduzido por Zhang et al. (2019) comparou as três plataformas em termos de sensibilidade, precisão, vício e custo. A Figura 1.4 resume os resultados divulgados pelos autores.

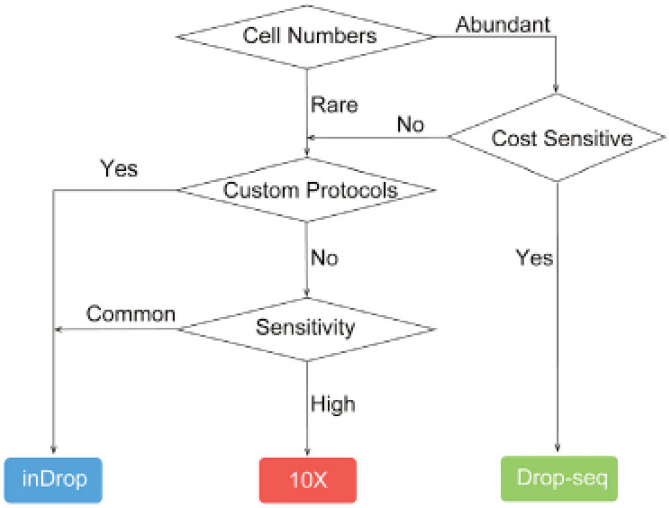


Figura 1.4 – Diagrama para escolha entre as plataformas Drop-seq, InDrop e 10X Genomics Chromium. Imagem retirada de Zhang et al. (2019).

As principais diferenças entres os protocolos experimentais ocorre nas etapas de captura da molécula, transcrição reversa, amplificação e preparação da biblioteca. De forma geral, a captura da molécula acontece via captura do poly(A)-tail RNA, embora determinados protocolos podem capturar todo RNA ou miRNAs (LAFZI et al., 2018). Após a lise celular, processo de dissolução celular causada pela ruptura da membrana plasmática, cadeias de oligonucleotídios se ligam a poly(A)-tail RNA. Este processo, captura mRNAs e exclui outros tipos, tais como rRNA e tRNA.

As moléculas de RNA são por natureza muito mais instáveis que o DNA, podendo sofrer degradação bastante acelerada. Além disso, a maioria das plataformas de sequenciamento utilizam o DNA como molécula de entrada. Por esses motivos, as moléculas de RNA capturadas passam pelo processo de transcrição reversa em moléculas de cDNA mais estáveis. Esta é uma etapa fundamental do experimento e a sensibilidade do sequenciamento esta diretamente ligada a eficiência da transcrição reversa (KOLODZIEJCZYK et al., 2015). Lafzi et al. (2018) apresentam uma discussão sobre as enzimas e aditivos específicos utilizados por diferentes protocolos para otimização da transcrição reversa.

Uma vez que as células isoladas apresentam uma escassa quantidade de material biológico é necessário realizar a amplificação do cDNA, que pode ser feita via reação em cadeia de polimerase (PCR) ou transcrição em vitro (IVT). Independente, do método utilizado a amplificação pode ocasionar distorções na proporção de genes quantificados. Uma estratégia desenvolvida por Islam et al. (2014) para correção dos vieses de amplificação é anexar UMIs para cada molécula de cDNA antes da amplificação. Cada UMI caracteriza unicamente as moléculas presentes na amostra, desse modo, após a amplificação a expressão genética é quantificada contando o número de UMIs distintos alinhados a cada característica

do genoma, ao invés de contar todas as leituras amplificadas da respectiva característica como usualmente é feito.

Outro aspecto que diferencia os protocolos de scRNA-seq é a estratégia para quantificação do transcriptoma, ou seja, os protocolos podem adotar a leitura completa do transcrito (*full-lenght*) ou pela contagem das extremidades 3' ou 5' do transcrito (*3'- or 5'-end transcript*) que permitem a incorporação de UMI.

Na preparação da biblioteca o cDNA amplificado é fragmentado em várias partes e adaptadores especializados são ligados as extremidades de cada parte. Segundo Lafzi et al. (2018) a fragmentação pode ser realizada enzimaticamente (tagmentase ou DNase), quimicamente (zin, KOAc ou MgOAc) ou através de forças mecânicas (ultrassom).

Recentemente, vários pesquisadores tem realizado estudos para comparação entre os métodos de sequenciamento disponíveis para scRNA-seq. As comparações são baseadas em métricas de qualidade tais como número de genes detectados por célula e a expressão genética obtida pelo método. O estudo conduzido por Ziegenhain et al. (2017) fornece uma comparação entre seis métodos: CEL-seq2, Drop-seq, MARS-seq, SCRBseq, Smart-seq e Smart-seq2. Os autores apontam, por exemplo, que o Drop-seq oferece a melhor solução custo-benefício para sequenciamento de 254 células a uma profundidade de 250.000 leituras. Outro estudo realizado por Svensson et al. (2017) comparou a sensibilidade e acurácia de 15 protocolos em diferentes estudos e tipos de célula com base em sua capacidade de quantificar a expressão de RNAs controles (*spike-in*) de concentração conhecida. Ding et al. (2019) comparam a performance de sete métodos sendo dois de baixo rendimento (Smart-seq e CEL-Seq) e cinco de alto rendimento (10x Chromium, Drop-seq, Seq-Well, inDrops e sci-RNA-seq) em três tipos de amostras – mistura de células humanas e de ratos, células humanas mononucleares do sangue periférico e córtex de ratos – resultando em aproximadamente 92.000 células. Os autores concluem que os métodos de baixo rendimento tiveram performance similares, embora o CEL-seq tenha sido mais afetado pela contaminação de leituras de outras células. Por outro lado, entre os métodos de alto rendimento, 10X Chromium foi o melhor. O estudo realizado por Mereu et al. (2020) avaliou sistematicamente 13 métodos para diferentes tipos e estados de células. Os métodos foram avaliados em termos da capacidade de recapitular a estrutura original das células. O grupo define uma pontuação geral baseada em várias características dos métodos, tais como a capacidade da detecção de genes e grupos, e concluem que os métodos Quartz-seq2 e 10X Chromium apresentaram as melhores performance gerais. A Tabela 1.1, adaptada de Lafzi et al. (2018), apresenta um resumo das características dos principais métodos de sequenciamento para scRNA.

Tabela 1.1 – Resumo das características dos principais métodos de sequenciamento de scRNA.

Método	Forma de captura	UMI	Amplificação do cDNA	Cobertura do transcrito	Fragmentação	Referência
Smart-seq	Plate	Não	PCR	full length	Tagmentação	Ramskold et al. (2012)
Smart-seq2	Plate	Não	PCR	full length	Tagmentação	Picelli et al. (2013)
STRT-seq	Plate	Sim	PCR	5' end	DNase I	Islam et al. (2012)
STRT-seq-2i	Nanowell	Sim	PCR	5' end	Tagmentação	Hochgerner et al. (2017)
SCRB-seq	Plate	Sim	PCR	3' end	Tagmentação	Soumillon et al. (2014)
mcSCRB-seq	Plate	Sim	PCR	3' end	Tagmentação	Bagnoli et al. (2018)
Quartz-seq	Plate	Não	PCR	full length	Ultrassom	Sasagawa et al. (2013)
Quartz-seq2	Plate	Sim	PCR	3' end	Ultrassom	Sasagawa et al. (2018)
CEL-seq	Plate	Não	IVT	3' end	KOAc, MgOAc	Hashimshony et al. (2012)
CEL-seq2	Plate	Sim	IVT	3' end	Random priming	Hashimshony et al. (2016)
MARS-seq	Plate	Sim	IVT	3' end	Zinc	Jaitin et al. (2014)
Seq-Well	Nanowell	Sim	PCR	3' end	Tagmentação	Gierahn et al. (2017)
inDrops	Droplets	Sim	IVT	3' end	KOAc, MgOAc	Klein et al. (2015)
Drop-seq	Droplets	Sim	PCR	3' end	Tagmentação	Macosko et al. (2015)
10x Chromium	Droplets	Sim	PCR	3' ou 5' end	Tagmentação	Zheng et al. (2017)

A companhia Ilumina é a principal fornecedora de plataformas para sequenciamento genético, sendo responsável pela produção de mais de 90% dos dados gerados no mundo (ILUMINA, 2017). As plataformas da Ilumina utilizam a abordagem de sequenciamento por síntese química (SBS). A Figura 1.5 ilustra as seguintes etapas do sequenciamento realizado por dispositivos da Ilumina.

A. *Library Preparation*: cDNA amplificado é fragmentado em várias partes e adaptadores especializados são ligados as extremidades de cada parte.

B. *Cluster Amplification*: os adaptadores são conectados à superfície de sequenciamento e cada fragmento é copiado várias vezes por meio de "*bridge amplification*" resultando em distintos grupos com cópias idênticas do mesmo fragmento.

C. *Sequencing*: nucleotídeos marcados com cores são lançados na superfície. A medida que cada nucleotídeo se liga a uma base complementar, para cada grupo de fragmentos, seu sinal é emito e reconhecido como uma imagem que identifica a base. Este processo é repetido para cada uma das quatro bases e todos os grupos de fragmentos em paralelo, gerando milhões de leituras de sequências com aproximadamente 125 a 300 bases cada uma de comprimento.

D. *Alignment*: métodos computacionais são utilizados para alinhar cada sequência de nucleotídeo lida ao seu respectivo genoma de referência. Diferenças entre a sequencia obtida e o genoma de referência podem ser identificadas.

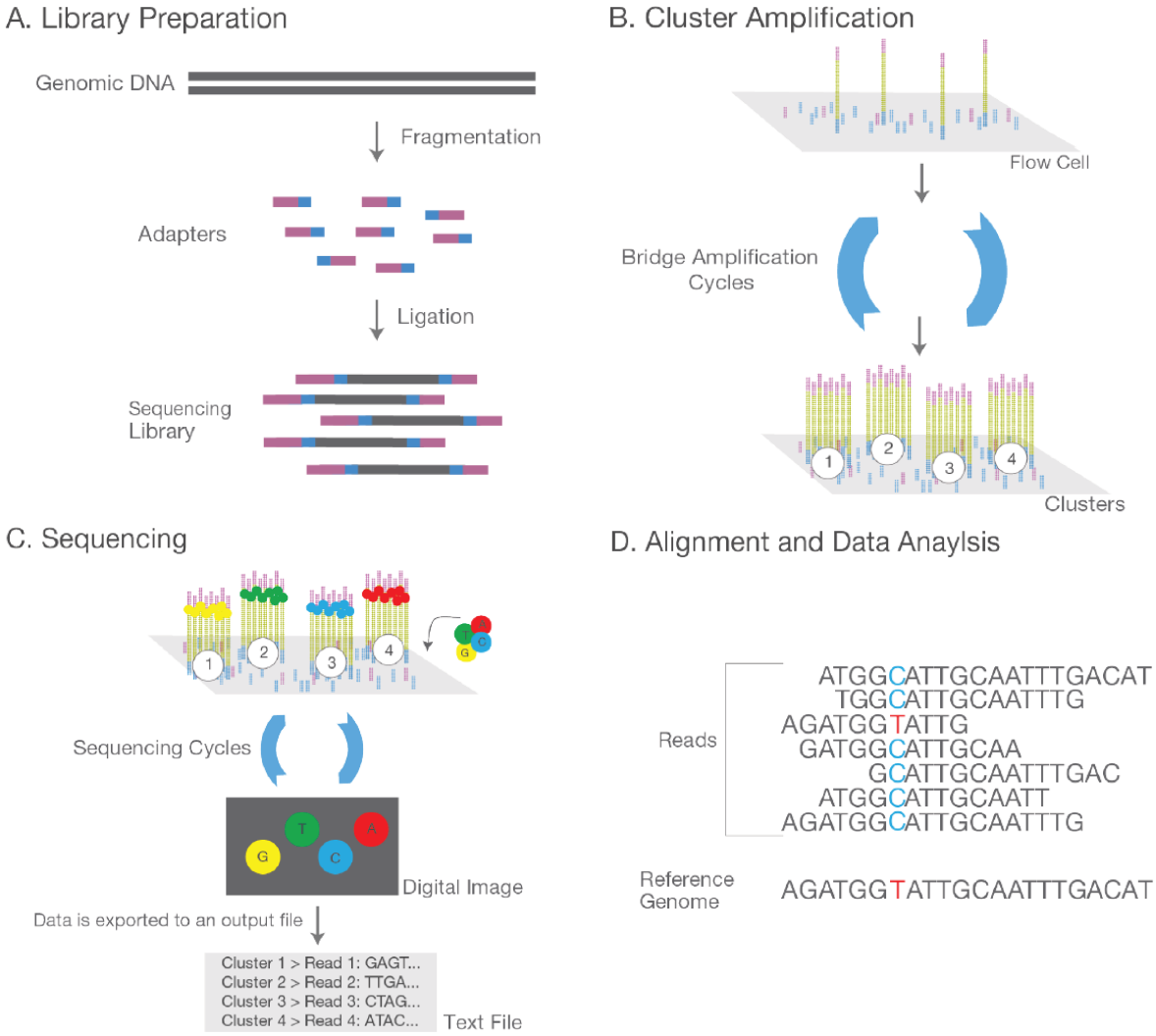


Figura 1.5 – Etapas utilizadas para sequenciamento em plataformas da Ilumina. Imagem adaptada de Ilumina (2017).

1.2.2 Pré-processamento dos dados de scRNA-seq

Tipicamente, os dados gerados por scRNA-seq são arquivos no formato FASTQ contendo as sequências de nucleotídeos encontradas. A etapa inicial para caracterização dos dados de scRNA-seq é o pré-processamento das sequências em uma matriz de contagem de expressão genética por *feature*, geralmente gene, e célula (AMEZQUITA et al., 2020). O pré-processamento consiste no mapeamento dos fragmentos sequenciados para um transcriptoma ou mesmo genoma de referência e a quantificação das expressões genéticas em uma matriz de contagem onde as linhas representam os genes e as colunas as células, como ilustrado na Tabela 1.2.

Tabela 1.2 – Representação genérica dos dados pré-processados provenientes de experimentos scRNA-seq.

	Célula 1	Célula 2	$\cdots$	Célula n
Gene 1	y_{11}	y_{12}	$\cdots$	y_{1n}
Gene 2	y_{21}	y_{22}	$\cdots$	y_{2n}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Gene g	y_{g1}	y_{g2}	$\cdots$	y_{gn}

Tendo em vista a complexidade e o volume dos dados obtidos em experimentos de scRNA-seq é fundamental a utilização de ferramentas computacionais que permitam acesso, gerenciamento e infraestrutura dos dados, juntamento com métodos especializados para a análise (AMEZQUITA et al., 2020). Seguramente, o projeto Bioconductor é a principal iniciativa mundial de código aberto que oferece soluções para análise de dados genômicos (HUBER et al., 2015). O Bioconductor é um repositório de pacotes desenvolvidas na linguagem R que além de oferecer uma documentação de alta qualidade faz uso de uma infraestrutura de dados comum com intuito de promover a interoperabilidade das bibliotecas (AMEZQUITA et al., 2020). Para esse fim, o Bioconductor utiliza um paradigma flexível e orientado a objetos S4 que permite o encapsulamento de vários componentes de objetos em uma única instância.

Em relação aos dados de scRNA-seq a classe de objeto *SingleCellExperiment* do pacote **SingleCellExperiment** (LUN; RISSO, 2019) oferece uma infraestrutura unificada para manipulação e análise de dados. Especificamente, esta classe de objeto é uma extensão da classe *SummarizedExperiment* (MORGAN et al., 2020) para scRNA-seq e utiliza uma estrutura de dados que armazena todos as características dos dados de scRNA-seq, especificamente, dados da expressão gênica, metadados por célula e reduções da dimensionalidade, além disso, fornece ferramentas para manipulações. A estrutura geral dos objetos da classe *SingleCellExperiment* é ilustrado na Figura 1.6.

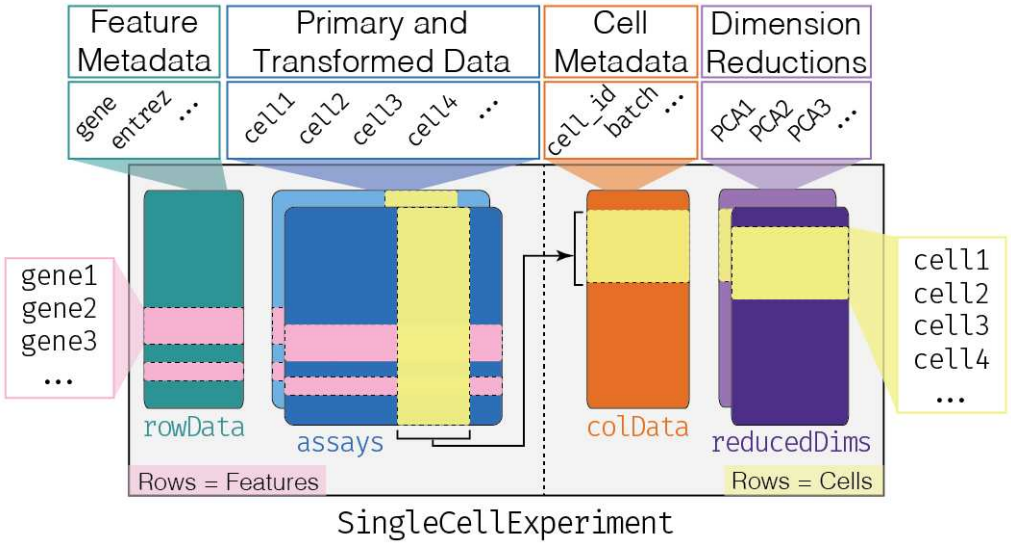


Figura 1.6 – Estrutura geral da classe de objeto `SingleCellExperiment`. Imagem retirada de Amezquita et al. (2020).

Diferentes métodos computacionais para pré-processamento dos dados de scRNA-seq tem sido proposto na literatura. Os pacotes do Bioconductor **scPipe** (TIAN et al., 2018) e **scruff** (WANG et al., 2019) oferecem soluções desenvolvidas em R. Além disso, os pacotes **DropletUtils** (LUN et al., 2019) e **tximetia** (LOVE et al., 2020) importam dados pré-processados por métodos desenvolvidos para linha de comando tais como Cell Ranger (10X Genomics) (ZHENG et al., 2017), Kallisto-Bustools (MELSTED et al., 2019) e Alevin (SRIVASTAVA et al., 2019).

Após a fase de pré-processamento para obtenção da matriz de contagem o fluxo de análise, conforme discutido por Luecken e Theis (2019) e Amezquita et al. (2020), segue as seguintes fases:

1. Processamento da matriz de contagem:
   (i) Métricas de controle de qualidade.
   (ii) Normalização.
   (iii) Seleção de *features*.
   (iv) Redução da dimensionalidade.
2. Análise estatística:
   (i) Agrupamento.
   (ii) Detecção de genes marcadores.
   (iii) Análise de anotação.
   (iv) Análise da expressão gênica.

(v) Análise de trajetória.

Os capítulos seguintes realizam uma revisão bibliográfica e análise crítica dos principais métodos analíticos para tratamento matriz de contagem e a metodologia estatística para análise dos dados.

2 Métodos de Processamento

Este capítulo discute as etapas usuais de processamento da matriz de contagem. Nesta etapa de análise, realiza-se o tratamento dos vieses do experimento e outras fontes técnicas de ruídos indesejadas. Este tratamento é realizado de modo a preservar variações biológicas do fenômeno em estudo. Na Seção 2.1, discussões para identificar e filtrar células e genes com baixa qualidade são apresentadas. A aplicação de métodos de normalização com intuito de mitigar efeitos técnicos é abordada na Seção 2.2. Para reduzir o custo computacional e eliminar genes não interessantes, métodos para seleção de *features* são discutidos na Seção 2.3. Este capítulo finaliza com a Seção 2.4 apresentando as motivações e definições das técnicas de redução de dimensionalidade.

2.1 Controle de Qualidade

Os métodos que serão discutidos neste seção para identificar células de baixa qualidade são baseados na matriz de contagem obtida na etapa de pré-processamento dos dados. Células com baixa qualidade resultam das etapas experimentais do scRNA-seq, discutidas no Capítulo 1. Especificamente, os seguintes problemas podem ocorrer, conforme a etapa experimental.

1. **Dissociação da célula:** múltiplas células podem ser isoladas de forma conjunta. Além disso, pode ocorrer dano celular durante a dissociação implicando em perda e degradação do RNA.

2. **Captura da célula:** dispositivos podem falhar na captura, duplicar células ou até mesmo capturar múltiplas células. Alguns métodos apresentam restrições em relação ao tamanho da célula, levando a vieses de seleção de tamanho, por exemplo, o protocolo Drop-seq. Também, é possível a introdução de vieses devido ao tipo de célula.

3. **Transcrição reversa:** muitos transcritos podem ser perdidos nesta etapa. Islam et al. (2014) estimam que somente 10% a 20% dos transcritos sofrem o processo de transcrição reversa.

4. **Amplificação:** os cDNAs transcritos podem ser amplificados mais de uma vez. Métodos que utilizam UMI buscam corrigir vieses de amplificação (LUO; ZHANG, 2018).

5. **Preparação da biblioteca e sequenciamento:** A multiplexação de amostras nem sempre é perfeita, sendo que o número de leituras por célula pode variar bastante.

Segundo Kolodziejczyk et al. (2015) as etapas de transcrição reversa e amplificação são as que principalmente contribuem para inserção de ruídos técnicos nos dados. Nesse sentido, vários protocolos tem sido refinados para melhorar a eficiência e cobertura dos transcritos e naturalmente reduzir ruídos técnicos dos experimentos. Em particular, o uso de UMIs e genes controle (*spike-in*) são as principais estratégias utilizadas (LUO; ZHANG, 2018).

Como discutido no Capítulo 1, os UMIs são sequências aleatórias de nucleotídeos que os protocolos do tipo *microfluidic droplet-based* utilizam para corrigir possíveis efeitos técnicos da amplificação, pois cada UMI caracteriza unicamente as moléculas, sendo utilizados para obter a matriz de contagem. Por outro lado, os genes controle (*spike-in*) são utilizados para quantificar variações técnicas que podem ocorrer durante o experimento. Isso ocorre pois, o número de genes controle adicionados em cada célula é conhecido, fornecendo uma referência as medidas empíricas obtidas para outros genes e portanto permitindo uma calibração quantitativa do ruído técnico (VALLEJOS; MARIONI; RICHARDSON, 2015). O conjunto de 92 moléculas extrínsecas derivadas do *External RNA Controls Consortium* (ERCC) introduzido por Jiang et al. (2011) tem sido a principal escolha de genes controle.

As células definidas como sendo de baixa qualidade geralmente são caracterizadas por baixa contagem, poucos genes expressos e alta proporção de genes controles ou mitocondriais. Tais células podem afetar negativamente a condução de análises posteriores, Amezquita et al. (2020) relatam, por exemplo, os seguintes problemas:

- Formam seus próprios grupos de células, gerando complicações nas interpretações. O caso mais óbvio é causado por proporções de genes mitocondriais em determinadas células, isso é consequência direta do dano celular.
- Distorcem a real heterogeneidade entre as células, principalmente na estimação de componentes principais. As primeiras componentes capturam diferenças na qualidade da célula ao invés de diferenças biológicas.
- Podem conter *features* (genes ou transcritos) altamente desregulados levando a problemas na normalização gênica.

Nesse sentido, com intuito de ao menos mitigar efeitos técnicos inerente ao experimento, essas células e/ou *features* devem ser identificados e, em algumas situações, removidas do estudo.

2.1.1 Nível da Célula

Considerando as *features* como sendo moléculas de genes, as principais métricas para controle de qualidade das células são:

- **Tamanho da biblioteca:** soma das contagens ao longo dos genes. Em geral, células com tamanho pequeno de biblioteca são de baixa qualidade, devido, por exemplo, perda do RNA durante a preparação da biblioteca.

- **Total de genes detectados:** número total de genes endógenos com contagem diferente de zero. Células com poucos genes expressos provavelmente são de baixa qualidade, pois os transcritos podem não ter sido corretamente capturados.

- **Proporção de genes controles (*spike-in*):** razão entre total de genes controles detectados e total de genes detectados (incluindo os controles). Como a mesma quantidade de RNA controles foi adicionada a cada célula, assim, células com alta proporção de genes controles é um indicativo que os genes endógenos foram perdidos em algum momento do experimento.

- **Proporção de genes mitocondriais:** razão entre total de genes detectados na mitocôndria do genoma e total de genes detectados. Altas proporções são indicativos de células com baixa qualidade, possivelmente por causa da perda de RNA citoplasmático das determinadas células (ISLAM et al., 2014).

A identificação das células com baixa qualidade é realizada estudando o comportamento da distribuição das métricas calculadas. Este comportamento pode ser explorado de forma univariada, isto é, para cada métrica, independentemente, utiliza-se técnicas que determinam limiares definindo células com baixa qualidade. Por outro lado, pode-se encontrar células atípicas considerando o comportamento conjunto das métricas. Em ambas as abordagens, supõe-se que: (i) as métricas de qualidade são independentes da natureza biológica das células e (ii) a maioria das células são de alta qualidade.

2.1.1.1 Abordagens Univariada

O método mais simples desta classe consiste em subjetivamente definir valores limites para cada uma das métricas de qualidade. Por exemplo, pode-se considerar células com baixa qualidade se o tamanho da biblioteca for menor que 100.000 leituras, o total de genes detectados é menor que 5.000, a proporção de genes *spike-in* é menor que 10% ou a proporção de genes mitocondriais é menor que 10% (AMEZQUITA et al., 2020).

Embora simples, esta estratégia requer sensibilidade técnica acerca da natureza biológica das células e dos procedimentos experimentais adotados, uma vez que, o mesmo protocolo e tipo de célula pode produzir matrizes de contagens distintas devido as peculiaridades das etapas experimentais. Nesse sentido, a técnica mais utiliza consiste em obter limites empíricos das métricas calculadas através do desvio absoluto mediano (MAD). O MAD é um estimador robusto para a escala de uma distribuição. Formalmente, seja $X_1, \ldots, X_n$ uma sequência variáveis aleatórias independentes e identicamente distribuídas,

então

$$\text{MAD}_n = b\,\text{med}_i\left|X_i - \text{med}_j(X_j)\right|, \quad i,j = 1,\ldots,n \tag{2.1}$$

em que $\text{med}_i(X_i)$ é um estimador para a mediana de X_i e b é uma constante que depende da distribuição de X_i para que MAD_n seja consistente.

É comum supor que as métricas de qualidade ou transformações das mesmas, por exemplo, log do tamanho da biblioteca, seguem uma distribuição normal com média μ e variância σ^2. Neste caso, têm-se que $b = 1/\Phi^{-1}(0.75)$, em que $\Phi(\cdot)$ é a função de distribuição da Normal padrão. Portanto, $\mathbb{E}(\text{MAD}_n) = \sigma$, quando $n \to \infty$, isto é, o MAD é um estimador consistente para o desvio-padrão (escala) (HUBER, 1981).

Na prática, a célula $i = 1,\ldots,n$ é considerada de baixa qualidade conforme a métrica j, se

$$x_{ij} > \text{med}_i(x_{ij}) + k \cdot \text{MAD}_{jn} \quad \text{ou} \quad x_{ij} < \text{med}_i(x_{ij}) - k \cdot \text{MAD}_{jn} \tag{2.2}$$

em que x_{ij} é o valor observado da métrica j na célula i, $\text{med}_i(x_{ij})$ é a estimativa da mediana da métrica j, MAD_{jn} é a estimativa do MAD para a métrica j e k é uma constante conhecida que controla ponto de corte.

Em geral, considera-se $k = 3$, pois este filtro mantém 99% dos valores não extremos quando a distribuição é normal (HUBER, 1981). O critério (2.2) é aplicado para todas as métricas, e finalmente define-se as células de baixa qualidade aquelas que foram classificadas em pelo menos uma das métricas adotadas.

2.1.1.2 Abordagem Multivariada

Nesta abordagem os valores atípicos são identificadas em um espaço de alta dimensão com base nas métricas para cada célula. Em particular, o pacote **scater** (MCCARTHY et al., 2017) do Bioconductor utiliza o método proposto por Hubert e Van der Veeken (2008) e implementado no pacote **robustbase** (MAECHLER et al., 2020).

O método proposto por Hubert e Van der Veeken (2008) é uma extensão multivariada da técnica *adjusted outlyingness* (AO) introduzida por Hubert e Vandervieren (2008) para obter ajustes das barras de limites do gráfico Boxplot em distribuições assimétricas.

Seja $\boldsymbol{x} = (x_1,\ldots,x_n)$ uma amostra aleatória, no caso univariado a medida AO é dada por

$$\text{AO}^{(1)}(x_i,\boldsymbol{x}) = \begin{cases} \dfrac{x_i - \text{med}(\boldsymbol{x})}{w_2 - \text{med}(\boldsymbol{x})} & \text{se, } x_i > \text{med}(\boldsymbol{x}) \\ \dfrac{\text{med}(\boldsymbol{x}) - x_i}{w_1 - \text{med}(\boldsymbol{x})} & \text{se, } x_i < \text{med}(\boldsymbol{x}) \end{cases} \tag{2.3}$$

em que w_1 e w_2 são, respectivamente, os limites inferiores e superiores do Boxplot ajustado. Em particular, esses limites são definidos por

$$\begin{cases} w_1 = Q_1 - 1.5e^{-4\text{MC}}\text{IQR} \quad \text{e} \quad w_2 = Q_3 + 1.5e^{3\text{MC}}\text{IQR} \quad \text{se MC} > 0 \\ w_1 = Q_1 - 1.5e^{-3\text{MC}}\text{IQR} \quad \text{e} \quad w_2 = Q_3 + 1.5e^{4\text{MC}}\text{IQR} \quad \text{se MC} < 0 \end{cases} \tag{2.4}$$

em que Q_1 e Q_3 são, respectivamente, o primeiro e terceiro quartil, $IQR = Q_3 - Q_1$ é a amplitude interquartil. Além disso, MC (*medcouple*) é uma medida robusta de assimetria introduzida por Brys, Hubert e Struyf (2004) e definida como

$$\text{MC} = \underset{x_i \leqslant Q_2 \leqslant x_j}{\text{med}} h(x_i, x_j) \tag{2.5}$$

sendo Q_2 a mediana amostral e h a função Kernel definida por

$$h(x_i, x_j) = \frac{(x_j - Q_2) - (Q_2 - x_i)}{x_j - x_i}.$$

Considere que, agora, a matriz $n \times p$ dada por $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ em que $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$, representando as p métricas de qualidade das n células consideradas. A versão multivariada da medida AO proposta por Hubert e Van der Veeken (2008) para o vetor de observações $\boldsymbol{x}_i$ é definida por

$$\text{AO}_i = \text{AO}(\boldsymbol{x}_i, \mathbf{X}) = \sup_{\boldsymbol{a} \in \mathbb{R}^p} \text{AO}^{(1)}(\boldsymbol{a}^\top \boldsymbol{x}_i, \mathbf{X}\boldsymbol{a}) \tag{2.6}$$

em que $\boldsymbol{a}$ é um vetor de direção e $\boldsymbol{a}^\top \boldsymbol{x}_i$ pode ser interpretado como a projeção do vetor $\boldsymbol{x}_i$ na direção $\boldsymbol{a}$. Conforme apontam os autores, na prática o AO não pode ser calculado projetando as observações em todas as direções. Nesse sentido, os autores propõem considerar $m = 250p$ direções.

Uma vez que o AO multivariado é calculado para cada observação utiliza-se essa informação para identificar células de baixa qualidade. Em particular, a estratégia empregada é aplicar a regra do MAD descrita em (2.2) considerando os AO_i de cada célula.

2.1.2 Nível da *Feature*

Embora menos comum, métricas de qualidade também podem ser aplicadas ao observar o comportamento da outra dimensão da matriz de contagens, isto é, as *features* (genes ou transcritos). Por exemplo, *features* com contagem igual ou próxima a zero são problemáticas, pois podem interferir e prejudicar análises subsequentes. Por este motivo, as métricas habitualmente empregadas nestas análises são níveis médios de expressão e proporção de presença (do gene ou transcrito) ao longo das células (MCCARTHY et al., 2017).

Lun, McCarthy e Marioni (2016) sugerem utilizar o nível de expressão da *feature* para filtragem, uma vez que uma *feature* será mantida desde que tenha expressão suficiente

em qualquer subconjunto das células. Sendo assim, os genes ou transcritos presentes em poucas células exigem níveis de expressão mais altos nessas células, o que não é indesejável e evita a seleção da genes não informativos, com baixa expressão em poucas células, que contribuem pouco para análises subsequentes. Em contrapartida, considerando a proporção de presença do gene ou transcrito ao longo das células pode resultar na não detecção de subpopulações raras de células.

2.1.3 Fatores Experimentais

Estudos mais complexos de scRNA-seq consideram amostras sob diferentes condições experimentais, por exemplo, distintas profundidades de sequenciamento, diferentes quantidades de RNA controle (*spike-in*) adicionado ou até mesmo indivíduos em diferentes estados de saúde (saudáveis versus doentes). Naturalmente, isso resulta em diferenças sistemáticas na expressão gênica observada entre as amostras, denominado efeito *batch* (AMEZQUITA et al., 2020). Tais efeitos são problemático, pois podem ser os principais fatores de heterogeneidade nos dados, mascarando as diferenças biológicas relevantes e complicando a interpretação dos resultados.

Nestes casos, a análise de todos os tipos de células em conjunto compromete o desempenho das métricas de qualidade, podendo levar à exclusão total de um tipo de célula, uma vez que a suposição fundamental das métricas é que são provenientes da mesma distribuição. A estratégia para evitar tais complicações é conduzir a análise de controle de qualidade separada conforme os efeitos sistemáticos existentes no estudo. Por exemplo, calcular diferentes MAD para cada *batch*.

2.1.4 Gráficos de Diagnósticos

Uma estratégia bastante utilizada na inspeção da qualidade das células é analisar graficamente a distribuição empírica das métricas. Para ilustrar os principais gráficos de diagnósticos utilizados no controle de qualidade, o conjunto de dados do estudo de Zeisel et al. (2015) será utilizado. Neste experimento, os autores investigaram dois tecidos do cérebro de ratos: o córtex somatossensorial e o hipocampo. As células individuais foram isoladas usando o sistema de microfluido Fluidigm C1 e a preparação da biblioteca foi realizada em cada célula usando um protocolo baseado em UMI. Genes controles (*spike-in*) do tipo ERCC bem como genes mitocondriais foram considerados no experimento de modo a avaliar diferenças gênicas sistemáticas. Após o pré-processamento das amostras sequenciadas a matriz de contagens resultou em 3005 células e 19839 genes.

A abordagem baseada no MAD foi aplicada considerando como fator experimental o tipo de tecido, dessa forma, diferentes valores para a mediana e o MAD foram calculados para as métricas conforme o tipo de tecido. Finalmente, uma célula foi classifi-

cada como sendo de baixa qualidade seguindo o critério (2.2) para cada uma das métricas. A distribuição de cada métrica de qualidade conforme o tecido do cérebro é apresentada na Figura 2.1. Ao todo 188 células foram consideradas de baixa qualidade, sendo 9 devido ao tamanho da biblioteca, 55 devido a quantidade de genes detectados, 62 por apresentarem alta concentração de genes controles e 110 por apresentarem alta concentração de genes mitocondriais. Naturalmente, as células de baixa qualidade, estão em geral, concentradas nas caudas das distribuições das células.

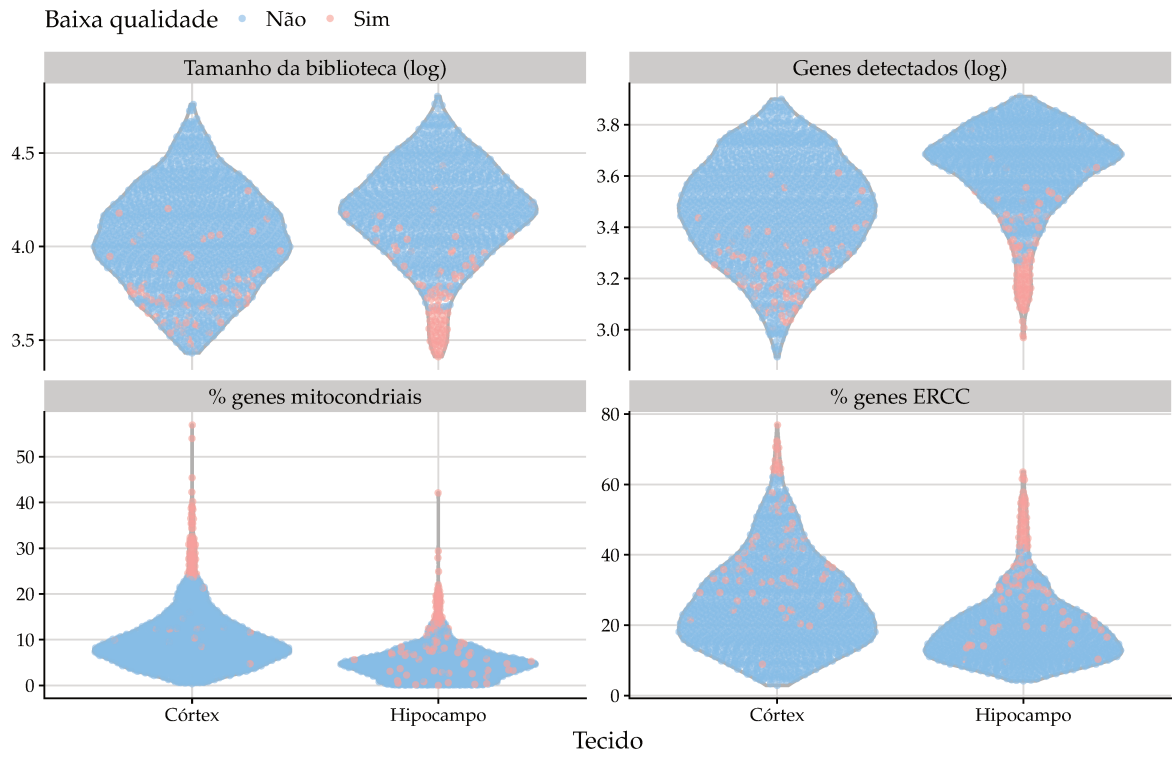


Figura 2.1 – Distribuições das métricas de qualidade conforme tecido para o conjunto de dados de Zeisel et al. (2015). Cada ponto representa uma célula e está colorido conforme o critério de baixa qualidade.

Outro gráfica de diagnóstico útil é verificar a relação entre genes mitocondriais e alguma outra métrica, por exemplo o tamanho da biblioteca. Objetivo deste gráfico é confirmar que não há células com altas contagens e alta proporção de genes mitocondriais, garantindo que células de alta qualidade que são metabolicamente ativas não sejam removidas (AMEZQUITA et al., 2020). Conforme a Figura 2.2, neste exemplo, não observa-se células no canto superior direito, que seria um indicativo de células metabolicamente ativas e não danificadas.

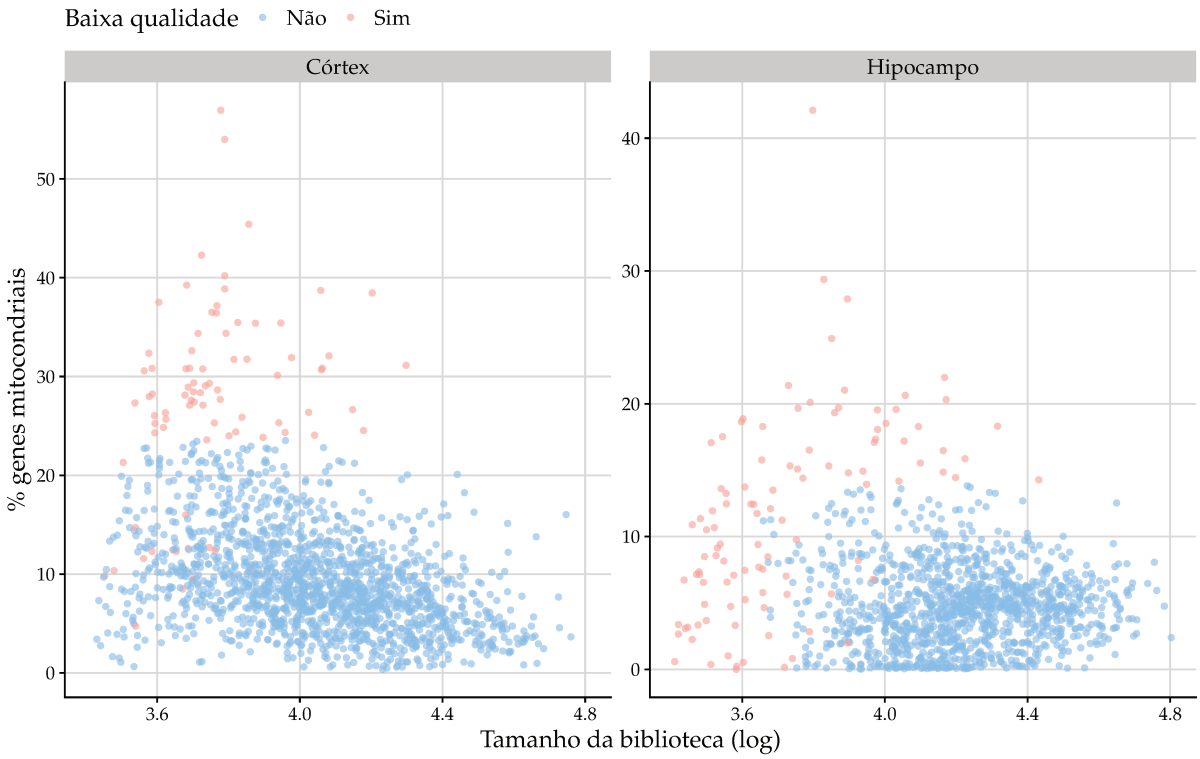


Figura 2.2 – Percentual de genes mitocondriais versus tamanho da biblioteca conforme tecido para o conjunto de dados de Zeisel et al. (2015). Cada ponto representa uma célula e está colorido conforme o critério de baixa qualidade.

Importante ressaltar que estudos empíricos demonstram que essas quatro métricas exibem fraca correlação umas com as outras, uma vez que avaliam diferentes critérios. Fato esse que motiva o uso simultâneo das métricas para de forma a capturar diferentes aspectos da qualidade técnica (AMEZQUITA et al., 2020).

2.2 Normalização

A normalização dos dados é uma etapa comum na análises de dados ômicos, em particular, em estudos que utilizam tecnologias de alto rendimento. As técnicas de normalização buscam minimizar o efeito de ruídos técnicos presentes nos dados, sem afetar o comportamento de fatores biológicos envolvidos. Em experimentos de scRNA-seq, os ruídos técnicos presentes nas amostras surgem, em geral, de diferenças técnicas na captura do cDNA ou na eficiência de amplificação de PCR entre as células. Tais diferenças são reveladas principalmente com o excesso de zeros presentes, isto é, para um dado gene (*feature*) muitas células não detectaram sua expressão.

2.2.1 Fatores de Escala

Seja X_{ig} a variável aleatória denotando a contagem de *reads* ou UMIs da célula i no gene g. De forma geral, os métodos de normalização discutidos nesta seção consistem em obter um fator de escala, s_i, que representa um viés especifico da célula i, assim, a expressão normalizada é definida por:

$$Y_{ig} = \frac{X_{ig}}{s_i}. \tag{2.7}$$

Ou seja, os métodos diferem na forma de estimar os fatores de escala (s_i) das células. A abordagem mais simples considera que s_i é o tamanho da biblioteca padronizado de tal forma que $\mathbb{E}(s_i) = 1$, isto é, s_i tem média igual a um. A principal desvantagem dessa abordagem é a suposição de que não há desequilíbrio na contagem dos genes diferencialmente expressos entre qualquer par de células. Na prática, qualquer regulação positiva para um subconjunto de genes é cancelada pela mesma magnitude de regulação negativa em um subconjunto diferente de genes (AMEZQUITA et al., 2020). Isso significa que na prática este tipo de normalização não é adequado, uma vez que a maioria de problemas reais existe uma diferença gênica presente.

Uma tentativa de contornar esta restrita suposição é a normalização baseada em genes controle (*spike-in*). Este método baseia-se na suposição de que de que a mesma quantidade de RNA controle (*spike-in*) foi adicionada a cada célula e respondem a vieses da mesma maneira relativa que os genes não controles. Para remover os vieses experimentais, o fator de escala, s_i, é o tamanho da biblioteca para os genes controles. Conforme Lun et al. (2017) discutem, existem dois pontos críticos que afetam o uso desta abordagem. O primeiro é a dificuldade de adicionar de forma consistente a mesma quantidade de RNA controle em cada célula. O segundo é que os transcritos do RNA controle podem não se comportar da mesma forma que os outro transcritos.

2.2.2 Método *Deconvolution*

Conforme discutido, uma característica peculiar dos dados gerados em scRNA-seq é o excesso de contagens iguais a zero. Motivado por isso e pelo desejo de obter um método geral que possa ser aplicado a qualquer conjuntos de dados Lun, Bach e Marioni (2016) introduziram o método de normalização por *deconvolution*, o qual tem sido bastante utilizado na literatura principalmente devido à sua disponibilidade no pacote **scran** do Bioconductor.

O objetivo deste método é aplicar a normalização nos valores somados das contagens a partir de um conjunto (*pool*) de células. Segundo Lun, Bach e Marioni (2016) as etapas do método *deconvolution* são:

1. Definir um *pool* de células;

2. Somar as contagens das células em cada *pool*;

3. Utilizando os valores somados normalizar o *pool* de células contra uma média de referência;

4. Repetir as etapas (1)-(3) para vários diferentes *pools* de células para construir um sistema linear;

5. A partir dos fatores de escala dos *pools* obter os fatores de escala para cada célula.

Seguindo a notação da Seção 2.2.1, para melhor compreensão das fases do método *deconvolution* seja X_{ij} uma variável aleatória que representa a contagem do gene i na célula j tal que $\mathbb{E}(X_{ij}) = \theta_j \lambda_{i0}$, sendo θ_j um viés especifico da célula e λ_{i0} a contagem esperada de transcritos i. Seja $Z_{ij} = X_{ij} t_j^{-1}$ em que t_j é uma constante, fator de escala para a célula j, assim, $\mathbb{E}(Z_{ij}) = \theta_j \lambda_{i0} t_j^{-1}$. Considere um *pool* k com um conjunto de células arbitrárias $\mathcal{S}_k$. Define-se V_{ik} como a soma dos Z_{ij} das células do conjunto $\mathcal{S}_k$ de tal forma que

$$\mathbb{E}(V_{ik}) = \lambda_{i0} \sum_{j \in \mathcal{S}_k} \theta_j t_j^{-1}.$$

Valores observados de V_{ik} em todos os genes representam a expressão do *pool* de células do conjunto $\mathcal{S}_k$. Considere, também, U_i como sendo a média de Z_{ij} em relação as N células do experimento, de tal forma que

$$\mathbb{E}(U_i) = \frac{\lambda_{i0}}{N} \sum_{j \in \mathcal{S}_0} \theta_j t_j^{-1}$$

em que $\mathcal{S}_0$ refere-se ao conjunto de todas as células.

Os valores observados de U_i representam a expressão média de uma pseudo-célula de referência. Segundo os autores o *pool* k de células é normalizado utilizando como referência os valores de U_i. Assim, define-se $R_{ik} = \dfrac{V_{ik}}{U_i}$ de tal forma que a esperança representa o verdadeiro fator de escala para o *pool* de células em $\mathcal{S}_k$, isto é,

$$\mathbb{E}(R_{ik}) \approx \frac{\mathbb{E}(V_{ik})}{\mathbb{E}(U_i)} = \frac{N \sum\limits_{j \in \mathcal{S}_k} \theta_j t_j^{-1}}{\sum\limits_{j \in \mathcal{S}_0} \theta_j t_j^{-1}} = \frac{\sum\limits_{j \in \mathcal{S}_k} \theta_j t_j^{-1}}{C} \tag{2.8}$$

em que C é uma constante que não depende do gene, da célula ou $\mathcal{S}_k$.

A aproximação presente em 2.8 assume que a variância de U_i é pequena devido à lei dos grandes números, o que de fato é razoável, uma vez que os dados de scRNA-seq possuem centenas de células. Denotando as realização das variáveis aleatórias X_{ij}, V_{ik}, U_i e R_{ik} como x_{ij}, v_{ik}, u_i e r_{ik}, respectivamente. O fator de escala para o *pool* de células em $\mathcal{S}_k$, $\mathbb{E}(R_{ik})$, é estimado considerando a mediana dos r_{ik} entre os genes, sob a suposição de que

a maioria dos genes não possuem diferença em suas expressões entre os *pools* e a média da pseudo-célula.

As estimativas de $\mathbb{E}(R_{ik})$ a partir de vários *pools* são utilizadas para obter uma estimativa de θ_j para cada célula. Para cada *pool* k, um equação linear é construída a partir de (2.8), substituindo $\mathbb{E}(R_{ik})$ pela sua estimativa e tratando $\theta_j t_j^{-1}$ para $j \in \mathcal{S}_k$ como parâmetros desconhecidos. A constante C pode ser ignorada considerando $C = 1$, pois não contribuí para as diferenças relativas entre fatores. A Figura 2.3 ilustra como o método *deconvolution* funciona.

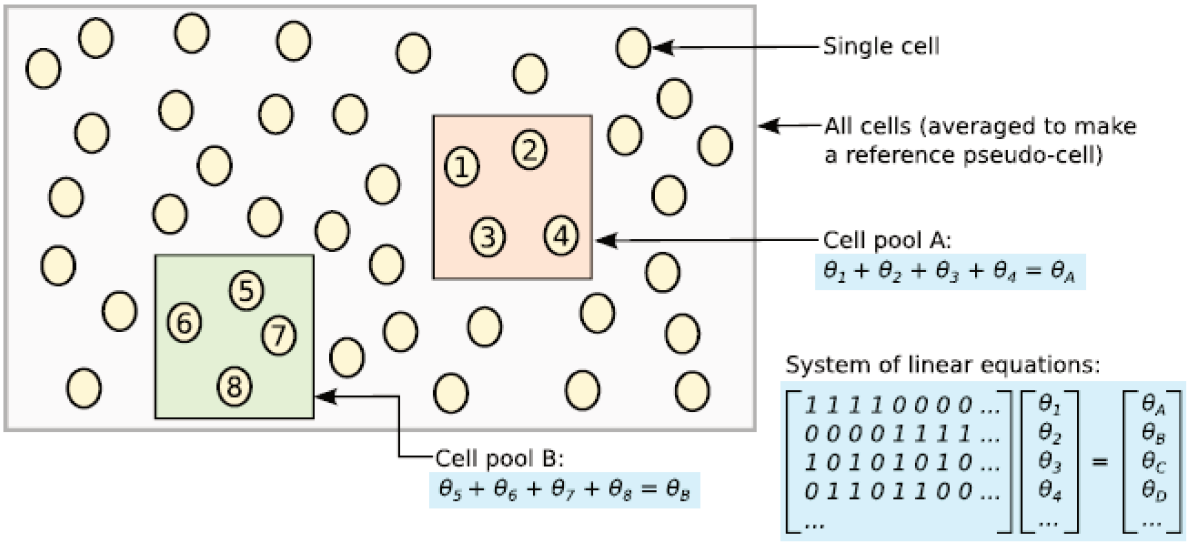


Figura 2.3 – Esquema do método *deconvolution*. A média de todas as células presentes no experimentos é considerada a pseudo-célula. Contagens para as células do *pool* A são somadas e normalizadas considerando como referência a pseudo-célula, resultando no fator de escala θ_A. Este fator é igual a soma dos fatores individuais das células, θ_j, $j = 1, \ldots, 4$, o qual é utilizado para formular a equação linear. Neste exemplo, assume-se que o termo t_j é igual a um. Repete-se este processo para vários *pool* (por exemplo, B) resulta na construção do sistema linear. Imagem retirada de Lun, Bach e Marioni (2016).

Segundo Lun, Bach e Marioni (2016) esta abordagem pode ser redundante, uma vez que, θ_j pode ser estimado diretamente a partir das contagens de cada célula. Todavia, a soma reduz o número de zeros que é o principal motivo de problemas dos métodos de fatores de escala discutidos na seção anterior.

2.2.3 Transformação logarítmica

No fluxo usual de análise, após obter as expressões normalizadas, aplica-se a transformação logarítmica. Seguindo a notação de (2.7), os valores normalizados na escala

log são definidos por:

$$Z_{ig} = \log(Y_{ig} + c) = \log\left(\frac{X_{ig}}{s_i} + c\right)$$

em que $c > 0$ é uma "pseudo contagem" que garante que a transformação esteja definida para todo $X_{ig} \geqslant 0$ e $s_i > 0$.

Essa transformação é motivada pela simplicidade computacional da função log e facilidade de interpretação. Em particular, essa transformação é útil pois garante que diferenças relativas nas expressões genicas das células sejam utilizadas em análises baseadas em distâncias, tais como métodos de redução de dimensionalidade e agrupamentos. A transformação logarítmica também reduz o impacto de flutuações estocásticas nas contagens de genes altamente expressos (LUN, 2018).

Apesar da popularidade, valores normalizados na escala logarítmica apresentam alguns problemas, tais como subestimação na estabilização da variância, que significa que a variabilidade de um gene é impulsionada mais por sua abundância do que por sua heterogeneidade biológica subjacente, e arbitrariedades na escolha da pseudo-contagem. Sob o ponto de vista, a normalização é aplicada para remover vieses entre amostras na escala de contagem, assim, para genes que não possuem diferenças em suas expressões a esperança das expressões gênicas é a mesma entre células. Entretanto, a esperança dos valores log-normalizados pode não ser a mesma, resultando em diferenças espúrias entre as células na escala log.

Lun (2018) mostra que a transformação logarítmica introduz vieses sistemáticos em genes que não possuem diferenças gênicas expressas. Para observar este fato, considere a aproximação de segunda-ordem de Taylor para a esperança de Z_{ig}, dada por:

$$\mathbb{E}(Z_{ig}) \approx \log(\mu_{ig} + c) - \frac{\mathrm{Var}(X_{ig})\, s_i^{-2}}{2\,(\mu_{ig} + c)^2}$$

em que $\mu_{ig} = \mathbb{E}\left(\frac{X_{ig}}{s_i}\right)$. O segundo termo da equação representa a discrepância entre a média da expressão genica dos valores log-normalizados e o log da média da expressão normalizada.[1]

Considere duas células $i = 1$ e $i = 2$ que diferem em seu fator de escala s_i. Assumindo que o gene g não possui diferença gênica entre as células, de forma que $\mathbb{E}\left(\frac{X_{ig}}{s_i}\right) = \mu_g$ para ambas as células. Então, a verdadeira mudança na escala log na esperança dos valores normalizados entre as células é dada por:

$$\delta_{12g} = \log\left(\frac{\mu_g}{\mu_g}\right) = 0.$$

[1] the mean of the log-normalized expression and the log-mean normalized expression.

Entretanto, na prática utiliza-se a diferença entre as esperanças de $Z_{ìg}$ como uma aproximação para δ_{12g}, dada por:

$$\Delta_{12g} \approx \frac{\mathrm{Var}(X_{2g})\, s_2^{-2} - \mathrm{Var}(X_{1g})\, s_1^{-2}}{2(\mu_g - 2)^2}. \tag{2.9}$$

Ou seja, em geral é diferente de zero, devido a diferenças nos fatores de escalas s_i e consequentemente nas variâncias, resultando em diferenças espúrias para genes que não possuem diferenças, isto é, $\delta_{12g} = 0$. Segundo Lun (2018) este efeito é aumentado quando aplica-se análises baseadas em distâncias. Por exemplo, considere a distância euclideana entre os valores de log-normalizados de duas células $i = 1$ e 2. Idealmente, a distância deveria ser zero, porém, conforme (2.9) a distância será aproximadamente

$$\sum_{g=1}^{J} \sqrt{\Delta_{12g}^2}.$$

Isso significa que a distância entre células com diferentes s_i será sistematicamente maior do que células com valores iguais de s_i. Como efeito prático na análise dos dados, grupos ou trajetórias de células espúrias identificados, respectivamente, pelas análises de agrupamento e trajetória, podem se formar apenas devido à transformação logarítmica.

Um fato que tem sido objeto de estudo na literatura de scRNA-seq é a inflação de zeros. Pesquisas recentes apontam que o excesso de zero é ocasionado pela transformação logarítmica, especialmente em protocolos que utilizam UMIs (HAFEMEISTER; SATIJA, 2019; TOWNES et al., 2019; SVENSSON, 2020). Entretanto, é importante ressaltar que sendo a quantificação gênica realizada no nível da célula é usual a presença de genes com contagem iguais a zero. Nesse sentido, as principais propostas na literatura buscam analisar os dados diretamente das contagens, não realizando a normalização e consequentemente transformação logarítmica nas contagens. Por exemplo, Townes et al. (2019) introduzem métodos para seleção de *features* e redução de dimensionalidade supondo que as contagens seguem uma distribuição multinomial. Enquanto que Hafemeister e Satija (2019), introduzem um método de normalização baseado nos resíduos de Pearson do modelo de regressão Binomial negativa com regularização.

2.3 Seleção de *Features*

Nesta etapa de análise, o objetivo é selecionar os genes que contribuam para explicar a natureza biológica do experimento e portanto reduzir a variabilidade presente que afeta a estrutura biológica do fenômeno bem como reduzir o número de *features* (genes) para melhor eficiência computacional das etapas posteriores. (AMEZQUITA et al., 2020).

Abordagens usuais para seleção dos genes consistem em assumir que diferenças biológicas entre os genes serão manifestadas com o aumento da variância. Uma vez que existe uma natural heterogeneidade entre os genes a variância não pode ser utilizada isoladamente para a seleção. Dessa forma, os métodos mais habitualmente utilizados propões diferentes abordagens para modelar a relação entre média-variância dos genes (YIP; SHAM; WANG, 2018). Recentemente, (TOWNES et al., 2019) propuseram uma abordagem assumindo que a contagem observada segue uma distribuição Multinomial, ou seja, os autores não utilizam os valore log-normalizados. Por outro lado, os autores sugerem utilizar a função *deviance* como uma forma de selecionar os genes com maiores variabilidades.

2.3.1 Relação Média-Variância

Conforme discutido na Seção 2.2.3, os valores log-normalizados não conseguem estabilizar a variância, isso significa que a variância de um gene é denominada mais por sua abundância do que por sua heterogeneidade biológica subjacente. Para levar em consideração este efeito, os métodos existentes exploram a relação entre média e variância, buscando estimar a variação de cada gene. Uma revisão extensiva desses métodos e comparação é realizada por Yip, Sham e Wang (2018).

Nesta seção descrevemos duas abordagens implementadas no pacote `scran` do Bioconductor, uma vez que, tem sido, seguramente, a principal ferramenta computacional utilizada para analisar dados de scRNA-seq. Ambas as abordagens assumem que grande parte da expressão genica observada nas células é dominada por ruídos técnicos, assim, a expressão gênica é decomposta entre uma componente biológica da célula e técnica devido ao experimento.

A primeira abordagem ajusta um modelo de tendência entre a média e a variância das expressões gênica log-normalizadas. A relação entre média e variância é descrita pela seguinte estrutura paramétrica

$$v_i = \frac{\alpha\, m_i}{m_i^{\delta} + \beta}, \quad i = 1, \ldots, n \tag{2.10}$$

em que v_i e m_i são a variância e média amostrais dos gene i e α, β e δ são parâmetros positivos e desconhecidos que são estimados via mínimos quadrados.

Na sequencia, considerando os valores ajustados, $\widehat{v}_i^2$, aplica-se regressão local ponderada, LOWESS, proposta por Cleveland e Devlin (1988), de tal forma a capturar quaisquer tendências restantes nos resíduos.

O ajuste obtido é utilizado para decompor a variância total de cada *feature* (gene) em dois componentes: (i) técnico, os valores ajustados e (ii) biológico, definido como a diferença entre a variação total e o componente técnico, isto é, o resíduo. O componente

biológico representa a variabilidade de interesse para cada gene e pode ser usado como métrica para a seleção das *features*.

A segunda abordagem disponível no pacote `scran` é a proposta de Brennecke et al. (2013), que sugerem estudar a relação entre o coeficiente de variação e a média. Em particular, os autores relacionam o parâmetro de dispersão da distribuição Binomial Negativa com o coeficiente de variação, propondo assim a seguinte relação

$$\mathrm{CV}_i^2 = \alpha + \frac{\beta}{m_i}, \quad i = 1, \ldots, n \tag{2.11}$$

em que CV_i denota o coeficiente de variação amostral do gene i e $\alpha, \beta > 0$ são parâmetros desconhecidos e estimados via mínimos quadrados.

O ajuste obtido do modelo (2.11) é utilizado para quantificar desvios da tendência em termos da razão entre o valor observado e ajustado, $\widehat{\mathrm{CV}}_i^2$. Ou seja, valores grandes dessa razão caracterizam genes com alta variabilidade e portanto são os mais interessantes de serem selecionados.

A partir das estimativas de variação de cada gene, seja diretamente pela variância ou coeficiente de variação, a próxima etapa é selecionar um subconjunto de genes com alta variação em sua expressão gênica. Naturalmente, um subconjunto maior reduz o risco de descartar sinais biológicos interessantes ao reter genes mais potencialmente relevantes, ao custo de aumentar o ruído de genes irrelevantes que podem mascarar o referido sinal. Neste sentido, Amezquita et al. (2020) discutem três usuais abordagens para seleção desse subconjunto, sendo elas

- Selecionar os k primeiros genes com as maiores variabilidades.
- Para cada *feature*, testar a hipótese nula de que a variância é igual a tendência. Assim, seleciona as *features* que obtiveram valor-p do teste abaixo de algum nível de significância especificado.
- Manter todas as *features* acima da variação estimada. Por exemplo, considerando a primeira abordagem selecionaríamos todos as *features* com componente biológico (resíduo) maior que zero. Por outro lado, considerando o CV, seleciona-se as *features* com razão acima de 1.

2.4 Redução da Dimensionalidade

Os dados de scRNA-seq são caracterizados pela alta dimensionalidade tanto no número de células (amostras) quanto na quantidade de *features* (genes). As técnicas de redução de dimensionalidade visam reduzir o número de dimensões dos dados de forma a manter a variabilidade intrínseca do fenômeno. Nos dados de scRNA-seq isso é possível ser

feito, pois é plausível assumir que diferentes genes estão correlacionados se forem afetados pelo mesmo processo biológico. Assim, não é necessários armazenar informações separadas para cada gene, mas em vez disso, pode-se compactar os genes em um número pequeno de dimensões. Utilizar os dados no formato compacto reduz o custo computacional em análises posteriores, como por exemplo, agrupamento, uma vez que os cálculos são realizados para um número muito menor do que para os milhares de genes. A redução da dimensionalidade permite também visualizar em dimensões menores padrões presentes no conjunto de dados.

Em outras palavras, as técnicas de redução de dimensionalidade buscam projetar as p *features* em um subespaço m-dimensional, tal que $m < p$. Essa projeção é realizada calculando m diferentes combinações (não)-lineares das *features*. Esta seção discute os principiais métodos de redução de dimensionalidade usualmente utilizados na análise de dados de scRNA-seq.

2.4.1 Análise de Componentes Principais

A análise de componentes principais (*principal component analysis* – PCA) é uma técnica multivariada que permite a redução de dimensionalidade da matriz $\mathbf{X}$ com dimensão $n \times p$ para um subespaço menor $n \times l$, com $l \leqslant p$. Considere que $\mathbf{\Sigma}$ represente a matriz de covariância de $\mathbf{X}$ sendo sua dimensão $p \times p$. Pelo teorema da decomposição espectral têm-se que a matriz de covariância pode ser escrita como

$$\mathbf{\Sigma} = \mathbf{E}\,\Lambda\,\mathbf{E}^{\top} \tag{2.12}$$

em que $\Lambda = \text{diag}(\boldsymbol{\lambda})$ é a matriz de autovalores $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)$ da matriz $\mathbf{X}$ e $\mathbf{E}$ é a matriz dos autovetores $\boldsymbol{e}_i, i = 1, \ldots, p$.

Dessa forma, as novas variáveis denominadas componentes principais são dadas por:

$$\boldsymbol{z}_i = \mathbf{X}\,\boldsymbol{e}_i, \qquad i = 1, \ldots, p. \tag{2.13}$$

Percebe-se então que são calculadas p componentes principais, isto é, uma para cada variável preditora. Todavia, na prática, retemos somente as m primeiras componentes que serão suficientes para explicar boa parte da variação total dos dados originais $\mathbf{X}$. Para evitar perda de informação, espera-se que a proporção de variância explicada pelas m primeiras componentes seja o mais próximo de um.

Sabendo que a proporção de variância explicada pela i-ésima componente é dada por $\lambda_i / \sum_{j=1}^{p} \lambda_j$. Então, o gráfico da proporção acumulada versus o número de componentes pode servir para determinar um número ideal de componentes que devem ser retidas.

2.4.2 t-SNE e UMAP

A redução da dimensionalidade também é utilizada para compactar os dados em 2 ou 3 dimensões e visualizar o comportamento das células. Como o PCA é uma técnica que busca combinação lineares apenas a variação ao longo de uma reta no espaço de alta dimensão é capturada por cada componente, portanto a visualização da estrutura dos dados utilizando as 2 componentes principais não é eficiente para capturar e visualizar os padrões dos dados. Este fenômeno é ilustrado na análise de dados conduzida no Capítulo 4.

O método *t-Stochastic Neighbor Embedding* (t-SNE) proposto por Van der Maaten e Hinton (2008) tem sido uma das principais técnicas de redução de dimensionalidade para visualização de dados de scRNA-seq. O t-SNE visa buscar uma representação de baixa dimensão dos dados que preserve as distâncias entre cada ponto e seus vizinhos no espaço de alta dimensão. Ao contrário do PCA, não se restringe a transformações lineares, assim, a representação em 2 componentes do t-SNE possui maior liberdade na forma de organizar cada amostra (célula) permitindo identificar padrões em populações complexas. Entre as desvantagens do t-SNE estão (i) é computacional mais intensivo que outros métodos e (ii) requer cuidadoso conhecimento sobre os parâmetros que controlam a forma da redução de dimensionalidade. Sobre os parâmetros do t-SNE é importante destacar que como envolve uma inicialização aleatório é recomendado (i) repetir a visualização várias vezes para garantir que os resultados são representativos e (ii) fixar uma semente para garantir a reprodutibilidade da análise. Além disso, é interessante inspecionar diferentes valores para o parâmetro de perplexidade o qual determina a granularidade da visualização. De forma geral, baixos valores favorecem a resolução de estruturas mais finas.

O método *Uniform Manifold Approximation and Projection* (UMAP) introduzido por McInnes, Healy e Melville (2018) é uma alternativa ao t-SNE para redução de dimensionalidade não linear. O UMAP é baseado na teoria da topologia algébrica e geométrica de Riemann. Comparado ao t-SNE o UMAP tende a construir grupos mais compactos com maior espaço vazio entre eles, isso é ilustrado na análise conduzida no Capítulo 5. Além disso, o UMAP preserva mais da estrutura global dos dados do que o t-SNE. Da perspectiva prática, o UMAP tem custo computacional muito menor do que o t-SNE. Em relação aos parâmetros que controlam o UMAP, o número de vizinhos e a distância miníma entre os pontos têm o maior efeito na granularidade da visualização do UMAP (MCINNES; HEALY; MELVILLE, 2018).

3 Análise Estatística

Embora seja dependente do objetivo relacionado ao estudo nesta etapa de análise, os dados tratados são utilizados para descrever e/ou revelar a natureza biológica subjacente ao fenômeno. As análises podem ser conduzidas no nível das células ou genes. Por exemplo, a identificação de grupos de células com perfis de expressão gênica similares é um tipo de análise no nível celular. Em contrapartida, é bastante comum verificar se existe diferença entre genes sob distintas condições experimentais, por exemplo, quais são os genes que tem expressões alteradas (aumentadas ou diminuídas) quando comparamos situações inflamatórias versus saudáveis.

Este capítulo aborda os métodos estatísticos utilizados na análise dos dados de scRNA-seq. Especificamente, uma discussão sobre métodos de agrupamentos, técnicas para seleção do número de grupos e avaliação da estabilidade dos grupos é apresentada na Seção 3.1. Em seguida, a Seção 3.2 apresenta uma discussão dos principais testes estatísticos de comparações múltiplas para detecção de genes marcadores em cada grupo encontrado na análise de agrupamento. Os principais métodos e pacotes do Bioconductor para realizar a anotação dos tipos de células predominantes em cada grupos são apresentados na Seção 3.3. Por fim, uma revisão dos modelos estatísticos para análise da expressão gênica é abordada na Seção 3.4.

3.1 Métodos de agrupamento

A análise de agrupamento (*cluster*) refere-se a um amplo conjunto de técnicas que visam encontrar padrões em um conjunto de dados agrupando as observações em grupos (*clusters*) de tal forma que observações dentro de cada grupo são mais similares em relação a observações pertencentes a diferentes grupos. Embora considerada uma das áreas de análise multivariada, conforme discute os seguintes autores Everitt e Dunn (2001), Rencher (2002), Johnson e Wichern (2007) e Everitt e Hothorn (2011), com o advento dos computadores e a era da informação a análise de agrupamento tornou-se um dos principais métodos da mineração de dados (*data mining*) para desvendar padrões em dados multidimensionais. Nesse sentido, como pode ser visto recentemente em Hastie, Tibshirani e Friedman (2009), James et al. (2013) e Kassambara (2017), é comum também na literatura referir a análise de agrupamento como um problema não supervisionado de aprendizado de máquina (*unsupervised machine learning*).

Existem diversos métodos de agrupamento disponíveis na literatura, que podem ser classificados em diferentes tipos (JAIN; MURTY; FLYNN, 1999). Seguramente, a distinção mais comum entre os tipos de métodos é dada pelos *clusters* hierárquicos e

não-hierárquicos. Segundo Rencher (2002), os métodos hierárquicos não exigem um número inicial de grupos e buscam construir uma hierarquia de grupos. Em contrapartida, os métodos não-hierárquicos são caracterizados pela necessidade de definir um número inicial de grupos, e também pela sua flexibilidade, uma vez que, as observações podem mudar de grupos.

No contexto de scRNA-seq a análise de agrupamento é utilizada para definir empiricamente grupos de células com perfis de expressão gênica semelhantes. O principal objetivo é resumir os dados em um formato digerível para interpretação humana, permitindo descrever a heterogeneidade da população em termos de rótulos que são facilmente compreendidos, ao invés de tentar compreender a variedade de alta dimensão na qual as células realmente residem. Após a definição dos grupos, é usual realizar a anotação biológica dos grupos identificando os tipos de células mais predominantes e/ou com base em genes marcadores caracterizar funções biológicas especificas de cada grupo. Assim, os grupos encontrados podem ser tratados como aproximações para conceitos biológicos mais abstratos, como tipos de células ou estados. Kiselev, Andrews e Hemberg (2019) apresentam e discutem diferentes técnicas utilizadas para agrupamento no contexto de *scRNA-seq*, entre os quais estão inclusos k-*means*, métodos de agrupamento hierárquico e métodos baseados em grafos.

Para agrupar as células em grupos é necessário considerar alguma medida de similaridade entre os pares de observações (células), sendo que em muitos casos ela é mensurada por alguma medida de distância. Quando a natureza das variáveis é contínua, usualmente utiliza-se a distância euclidiana (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Considere dois vetores p-dimensionais $\mathbf{x} = (x_1, \ldots, x_p)^\top$ e $\mathbf{y} = (y_1, \ldots, y_p)^\top$ resultantes da mensuração de p variáveis em cada um das n observações. A distância euclidiana entre $\mathbf{x}$ e $\mathbf{y}$ é definida como:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{j=1}^{p} (\mathbf{x}_j - \mathbf{y}_j)^2}. \tag{3.1}$$

Outros tipos de distâncias podem ser utilizada dependendo do contexto da análise bem como da natureza das variáveis. Recentemente, o trabalho de Shirkhorshidi, Aghabozorgi e Wah (2015) avaliou em diferentes conjuntos de dados e métodos de agrupamento 12 distâncias para variáveis continuas.

3.1.1 K-means

Seguramente o agrupamento pela técnica *k-means* é a mais difundida na literatura, especialmente pela rapidez computacional e fácil implementação comparada a outros métodos de agrupamento. No contexto da análise de dados provenientes de scRNA-seq o *k-means* agrupa as células em k distintos grupos. Cada célula é atribuída

ao grupo com o centroide mais próximo, isso é feito minimizando a soma dos quadrados dentro do grupo (*within-cluster sum of squares* – WCSS) e utilizando um chute inicial aleatório para os k primeiros centroides.

Dado um conjunto de N observações $\mathbf{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N)$, em que cada observação é um vetor d-dimensional. Para os dados de scRNA-seq, em geral, utiliza-se as P primeiras componentes principais que fornecem retem alta proporção da variabilidade das *features*. O método *k-means* busca particionar as N observações em $k \leqslant N$ conjuntos $S = \{S_1, \ldots, S_k\}$ que miniminizam a WCSS, isto é,

$$\arg\min_{\mathbf{S}} \sum_{c=1}^{k} \sum_{\mathbf{y} \in S_c} \|\mathbf{y} - \boldsymbol{\mu}_c\|^2 \tag{3.2}$$

em que $\boldsymbol{\mu}_c$ é o centroide das observações em S_c.

Para resolver o problema de otimização do *k-means* vários algoritmos foram propostos, sendo o mais comum o desenvolvido por Lloyd (1982). Dado um conjunto inicial de k centroides $\mu_1^{(1)}, \ldots, \mu_k^{(1)}$ para $t > 1$, o algoritmo de Lloyd (1982) consiste nas seguintes etapas:

1. **Atribuição:** Atribua cada observação ao grupo com a média mais próxima, isto é, aquele com a menor distância euclidiana ao quadrado. Matematicamente, o conjunto S_i no passo t é dado por:

$$S_i^{(t)} = \left\{ y_p : \left\| y_p - \mu_i^{(t)} \right\|^2 \leqslant \left\| y_p - \mu_j^{(t)} \right\|^2 \ \forall\, j, 1 \leqslant j \leqslant k \right\}$$

   em que cada y_p é atribuído a exatamente um dos conjuntos $\mathbf{S}^{(t)}$.

2. **Atualização:** Recalcula as médias (centroides) considerando as novas observações atribuídas a cada grupo. O centroide do grupo i no passo $t+1$ é dado por:

$$\mu_i^{(t+1)} = \frac{1}{\left|S_i^{(t)}\right|} \sum_{y_j \in S_i^{(t)}} y_j.$$

   em que $\left|S_i^{(t)}\right|$ indica a quantidade de observações no conjunto $S_i^{(t)}$.

O algoritmo converge quando as atribuições não mudam mais, todavia, é importante mencionar que não há garantias de que o algoritmo de Floyd convirja, conforme apontam Hartigan e Wong (1979). Em particular, os mesmo autores propuseram uma algoritmo alternativo para melhorar a convergência do *k-means*, este algoritmo é o padrão implementado na função `kmeans` do pacote **stats** no software R.

3.1.2 Determinação do número de grupos

Um dos desafios presentes na análise de agrupamento é a escolha do número ótimo de grupos. Diversas métricas para encontrar o número ótimo de grupos foram propostos na literatura. Uma extensiva revisão de 30 índices é apresentada em Milligan e Cooper (1985). Idealmente, os grupos resultantes não devem ter somente boas propriedades estatística, mas também fornecer interpretações que sejam relevante no contexto do estudo. Uma métrica usual para auxiliar a escolher do número ótimo de grupos é soma de quadrados total dentro de cada grupo. Neste critério, a variância total dentro de cada grupo é minimizada buscando fornecer agrupamentos onde as observações possuem menor variabilidade possível.

Outra importante métrica é a estatística GAP introduzida por Tibshirani, Walther e Hastie (2002), que pode ser aplicada a qualquer método de agrupamento. A estatística GAP compara a medida de homogeneidade do grupo, especificamente a variação total dentro de cada grupo, para diferentes tamanhos (k), com seus respectivos valores esperados a partir de alguma distribuição de referencia, em geral, por uma distribuição uniforme.

Suponha que agrupamos os dados em k grupos, $C_1, \ldots, C_k$, com C_r denotando o índice de observações no grupo r. Considere que a soma dos pares de distâncias para todas as observações no grupo r seja dada por

$$D_r = \sum_{i,j \in C_r} d_{ij} \tag{3.3}$$

em que d_{ij} denota a distancia entre as observações i e j.

Definimos a variação total dentro de cada grupo como sendo

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r. \tag{3.4}$$

Conforme Tibshirani, Walther e Hastie (2002), a estimativa de k por meio da estatística GAP procede da seguinte forma.

1. agrupe os dados observados, considerando $k = 1, 2, \ldots K$ como o número total de grupos e para cada k obtenha a medida W_k definida em (3.4);

2. gere B conjuntos de dados de referência a partir de uma distribuição uniforme sobre o intervalo dos valores observados das variáveis;

3. Para cada B conjuntos de dados crie os grupos, e então obtenha a medida definida em (3.4), denotaremos por W^*_{kb} para $b = 1, 2, \ldots, B$. A estatística GAP é então, definida por

$$\text{GAP}(k) = \frac{1}{B} \sum_{b=1}^{B} \log W^*_{kb} - \log W_k, \qquad k = 1, \ldots, K. \tag{3.5}$$

4. Define-se $s_k = \mathrm{sd}_k \sqrt{1 + 1/B}$, em que

$$\mathrm{sd}_k = \sqrt{\frac{1}{B} \sum_{b=1}^{B} \left[\log W_{kb}^* - \bar{l}\right]^2}$$

é o desvio padrão de $\log W_{kb}^*$ e $\bar{l} = \sum_{b=1}^{B} \log W_{kb}^*$.

5. Por fim, o número ótimo de grupos estimado é dado por

$$\widehat{k} = \text{menor } k \text{ tal que } \mathrm{GAP}(k) \geqslant \mathrm{GAP}(k+1) - s_{k+1}.$$

3.1.3 Estabilidade dos grupos

Uma tarefa posterior ao agrupamento das células é avaliar sua qualidade. Para tal, o conceito de estabilidade dos agrupamentos é utilizado para determinar se os grupos são reproduzíveis em um experimento replicado. Em outras palavras, verificar se os grupo encontrados são estáveis a pertubações dos dados, isto é, pequenas mudanças nos dados não alteram as conclusões do estudo. A técnica de *Bootstrap* não paramétrico é uma usualmente utilizado para esse propósito, pois permite gerar amostras do mesmo experimento de forma simples, eficiente e sem custo adicional, apenas re-amostrando o conjunto de dados. Nos dados de scRNA-seq existem duas maneiras para obter amostras *Bootstrap*. Pode-se reamostrar com reposição os dados dos genes (*features*) ou células.

No primeiro caso, o procedimento se resume em reamostrar com reposição os genes, obter os agrupamentos e repetir esse procedimento um número B grande de vezes. Para duas células que foram originalmente agrupadas juntas, verifica-se a porcentagem de vezes que elas foram agrupadas nas B replicações *Bootstrap*. Isso fornece uma medida de reprodutibilidade para o método de agrupamento adotado, com valores maiores indicando que os agrupamentos são mais estáveis. Essa abordagem é atraente pois as identidades das células (amostras) não são alteradas, permitindo uma comparação direta de agrupamentos entre os conjuntos de dados replicados e originais. Todavia, realizar o *Bootstrap* na visão genes (*features*) requer uma série de problemáticas suposições, tais como:

- Os genes são independentes uns dos outros, condicionais à estrutura subjacente. Isso não é verdade devido às correlações entre genes co-regulados dentro dos grupos, o que superestima a confiança da subestrutura estocástica.
- Os genes são intercambiáveis entre si. Isso não é atraente em princípio, já que o conjunto de genes usados em cada experimento é fixo; uma "réplica" gerada por genes de reamostragem tem pouca semelhança teórica com uma réplica genuína gerada com os mesmos genes.

- Em geral, realiza-se a análise de agrupamento considerando as P primeiras componentes principais ao invés da expressão gênica observada, o que dificulta o processo de reamostragem.

No *Bootstrap* reamostrando diretamente das células (observações) é razoável supor que as células foram retiradas independentemente da mesma população. O procedimento adotado pelo pacote **bluster** (LUN, 2020) consiste em avaliar a estabilidade dos agrupamentos analisando a probabilidade de co-atribuição. As seguintes etapas são realizadas:

1. Obtém amostras *Bootstrap* a partir das células, agrupa o novo conjunto de dados e obtém um novo conjunto de grupos.
2. Atribui-se cada grupo original ao novo grupo (obtido pela amostras Bootstrap) mais próximo com base em seus perfis de expressão gênica média.
3. Para cada par de grupos originais, calcula-se o número de iterações de *Bootstrap* nas quais eles são atribuídos ao mesmo grupo obtido pela amostra *Bootstrap*.
4. Uma alta frequência de co-atribuições indica os pares de grupos não são estáveis entre si.

A frequência de co-atribuição entre dois pares de grupos pode ser interpretada como a probabilidade de dois grupos serem parecidos. Probabilidades altas indicam que os dois grupos não são estáveis um em relação ao outro. Isso é de certa forma útil, pois permite avaliar a confiabilidade da separação para diferentes grupos. Essa abordagem foi introduzida por Aaron Lun, pesquisador, autor e mantenedor de diversos pacotes do Bioconductor para análise de dados de scRNA-seq. Em particular, essa abordagem está disponível na função `bootstrapStability` do pacote **bluster** (LUN, 2020).

3.2 Detecção de genes marcadores

Para interpretar os grupos encontrados na análise de agrupamento uma estratégia usual é identificar os genes que conduzem a separação entre os grupos. Seguindo a metodologia discutida em Amezquita et al. (2020) o estado da arte para detecção de genes marcadores na análise de dados de scRNA-seq consiste em realizar testes de comparações múltiplas das expressões gênicas entre todos os pares de grupos e consolidar os resultados desses testes em uma lista de candidatos a genes marcadores para cada grupo. Os métodos descritos nas próximas subseções são baseados, principalmente no pacote **scran** (LUN; MCCARTHY; MARIONI, 2016), o qual tem sido a principal ferramentas de *software* livre para análise de dados de scRNA-seq.

3.2.1 Teste t de Welch

Entre os testes de comparações o mais usual é o teste-t de Welch, sendo as principais justificativas do seu uso (i) rapidez computacional, uma vez que envolve apenas o cálculo de médias e variâncias de todos os genes em cada grupo e (ii) facilidade de interpretação dos resultados, uma vez que as análise são realizadas utilizando a expressão gênica $\log_2$-normalizadas é possível interpretar os resultados em termos do *log-fold* indicando a intensidade da diferença entre os grupos para cada gene.

Considere que z_{gi} é expressão gênica de um determinado gene no grupo g na célula i. Sejam $z_{i1}, \ldots, z_{in_i}$ e $z_{j1}, \ldots, z_{jn_j}$ os valores observados das n_i e n_j células dos grupos i e j, respectivamente. Assumindo que as amostras são independentes, a estatística do teste-t de Welch é dada por:

$$t_W = \frac{\bar{z}_i - \bar{z}_j}{\sqrt{\frac{s_i^2}{n_i} + \frac{s_j^2}{n_j}}} \tag{3.6}$$

em que $\bar{z}_i$ e $\bar{z}_j$ são, as médias das expressões gênicas dos grupos i e j, respectivamente e s_i^2 e s_j^2 são as variâncias das expressões gênicas dos grupos i e j, respectivamente. Sob a hipótese nula $\mathcal{H}_0 : \mu_i = \mu_j$ ou $\mathcal{H}_0 : \mu_i \leqslant (\geqslant) \mu_j$ a estatística (3.6) tem distribuição t de Student com o seguinte grau de liberdade:

$$\nu \approx \frac{\left(\frac{s_i^2}{n_i} + \frac{s_j^2}{n_j}\right)^2}{\frac{s_i^4}{n_i^2\,(n_i - 1)} + \frac{s_j^4}{n_j^2\,(n_j - 1)}}. \tag{3.7}$$

3.2.2 Teste de Wilcoxon-Mann-Whitney

O teste da soma dos postos de Wilcoxon, também conhecido como teste de Wilcoxon-Mann-Whitney (WMW) introduzido por Mann e Whitney (1947) é uma alternativa não paramétrica para realizar comparações múltiplas. Sejam $x_1, \ldots, x_n$ e $y_1, \ldots, y_m$ os n e m valores da expressão gênica observada em dois grupos distintos. Assumindo que as amostras (células) dentro de cada grupo são independentes a estatística do teste de WMW é definida por:

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} S(x_i, y_j) \tag{3.8}$$

em que

$$S(x_i, y_j) = \begin{cases} 1, & \text{se } y_j < x_i \\ 0, & \text{se } y_j > x_i. \end{cases}$$

O teste de WMW assume que processo gerador das amostras em cada grupo é uma distribuição contínua, portanto a possibilidade de $x_i = x_j$ para alguma par i e j

não é admitida (GIBBONS; CHAKRABORTI, 2011). Uma definição usual da estatística WMW que admite empates é dada pela seguinte expressão

$$S_T(x_i, y_j) = \begin{cases} 1, & \text{se } y_j < x_i \\ \dfrac{1}{2}, & \text{se } y_j = x_i \\ 0, & \text{se } y_j > x_i \end{cases}$$

substituindo a expressão $S(x_i, y_j)$ por $S_T(x_i, y_j)$ na Equação (3.8).

Na identificação de genes marcadores, em geral, têm-se o interesse em testar se um determinado gene é possui uma regulação alta (baixa) entre os grupos, então a hipótese nula é formulada por $\mathcal{H}_0 : F_X(x) \geqslant (\leqslant) F_Y(x)$. Sob a hipótese nula especificada a estatística (3.8) tem distribuição assintótica Normal com média e variância definidas por

$$\mu_U = \frac{n\,m}{2} \quad \text{e} \quad \sigma_U^2 = \frac{n\,m\,(n+m+1)}{12}, \tag{3.9}$$

respectivamente.

Na presença de empates a variância corrigida da distribuição da estatística (3.8) é dada por

$$\sigma_{U_T}^2 = \frac{n\,m}{12}\left[n+m+1-\sum_{i=1}^{k}\frac{t_i^3-t_i}{(n+m)\,(n+m-1)}\right]$$

em que t_i é o total de amostras que compartilham do ranque i e k é o número de ranques distintos.

A estatística do teste fornece uma interpretação interessante do efeito de diferença entre os dois grupos. Em particular, a estatística de WMW avalia diretamente a separação entre as distribuições da expressão gênica de diferentes grupos, uma vez que, a estatística (3.8) é proporcional a AUC, isto é, a área sob à curva Característica de Operação do Receptor (*Receiver Operating Characteristic* – ROC). Tal relação é definida por

$$\text{AUC} = \frac{U}{n\,m}.$$

A prova entre da relação da estatística (3.8) e a AUC é pode ser encontrada em Mason e Graham (2002) e Gibbons e Chakraborti (2011). Nas comparações múltiplas entre os grupos encontrados a AUC mede a probabilidade de concordância de uma célula aleatória em um grupo ter uma expressão mais alta do que uma outra célula aleatória em outro grupo, assim, valores de AUC próximos de 0 ou 1 indicam que o gene separa perfeitamente os dois grupos. Portanto, a estatística do teste de WMW aborda diretamente uma característica de um gene marcador, por outro lado, o teste t de Welch avalia indiretamente tal característica por meio das diferenças entre as médias e variâncias.

3.2.3 Direção e magnitude do efeito dos testes

Na análise da expressão gênica, o que significa afirmar que um gene é expresso diferencialmente? Em termos estatísticos, um gene é expresso diferencialmente se seu nível de expressão muda sistematicamente entre dois grupos, independentemente de quão pequena seja a diferença. Por outro lado, na discussão científica, é provável que um gene seja considerado diferencialmente expresso apenas se seu nível de expressão mudar em uma quantidade válida. Portanto, existe uma desconexão entre os conceitos estatísticos e biológicos para detectar genes expressamente diferentes (MCCARTHY; SMYTH, 2009). Em muitos estudos os genes são aceitos como expressos diferencialmente se satisfazem simultaneamente os dois critérios mencionados. Essa discussão tem sido abordada por vários pesquisadores, especificamente, no contexto de dados de *microarray* McCarthy e Smyth (2009) apresentam a abordagem TREAT (*t-tests relative to a threshold*) que é uma extensão da estatística t moderada via Bayes empírico apresentada por Smyth (2004).

Para os testes t de Welch e Wilcoxon discutidos a avaliação da direção e magnitude da diferença entre dois grupos pode ser incorporada reescrevendo a hipótese nula em termos de testes de hipóteses unilaterais. Realizar testes de hipóteses unilaterais é vantajoso, pois permite encontrar genes que são regulados positivamente em cada grupo de interesse, o que usualmente possui maior apelo interpretativo. Seguindo a notação apresentada na Seção 3.2.1, no caso do teste t de Welch, suponha que o interesse seja testar se a magnitude do *log-fold change* é igual, maior ou menor a um valor especificado τ. Assim, as três hipóteses nulas formuladas são definidas, respectivamente, por:

- Bilateral: $\mathcal{H}_0 : \mu_i - \mu_j = \tau$, o verdadeiro log-*fold change* é $-\tau$ ou τ com probabilidades iguais. Assim, o valor-p bilateral é calculado por $2 \times P(|t_W| > \tau \mid \mathcal{H}_0)$.
- Unilateral à direita: $\mathcal{H}_0 : \mu_i - \mu_j \geqslant \tau$, o verdadeiro log-*fold change* é τ. O valor-p unilateral é calculado por $P(t_W > \tau \mid \mathcal{H}_0)$.
- Unilateral à esquerda: $\mathcal{H}_0 : \mu_i - \mu_j \leqslant \tau$, o verdadeiro log-*fold change* é τ. O valor-p unilateral é calculado por $P(t_W < \tau \mid \mathcal{H}_0)$.

sendo μ_i e μ_j a média das expressões log-normalizadas e t_W a estatística definida em (3.6).

3.2.4 Incorporando fatores de variação conhecida

Em alguns estudos, o sequenciamento das células estão sujeitos ao desenho experimental que pode apresentar fatores conhecidos que interferem na suposição de independência entre as amostras, por exemplo, sexo do indivíduo, efeitos de lote. Nessas situações, se os efeitos não forem incorporados na análise podem interferir na detecção dos

genes marcadores, particularmente inflacionando a variância dos genes dentro dos grupos ou alterando a expressão média caso exista diferenças entre os fatores dentro dos grupos.

Para incorporar os fatores de variação conhecidos nas comparações entre os grupos utilizando os testes t de Welch ou teste de Wilcoxon-Mann-Whitney Lun, McCarthy e Marioni (2016) propõem realizar, para cada gene, as comparações múltiplas entre os grupos separadamente em cada nível do fator conhecido. Então, para cada comparação múltipla os valores-p obtidos nos diferentes níveis do fator são combinados utilizando a versão ponderada do método Z-score de Stouffer (STOUFFER et al., 1949). A estratégia de Stouffer consiste em combinar os os valore-p na seguinte estatística

$$Z = \frac{\sum_{i=1}^{k} \omega_i Z_i}{\sqrt{\sum_{i=1}^{k} \omega_i^2}} \tag{3.10}$$

em que $Z_i = \Phi^{-1}(1-p_i)$ com p_i sendo o valor-p obtido no nível i do fator, $\Phi(\cdot)$ a distribuição acumulada da Normal padrão e e w_i os pesos de cada valor-p.

O peso para cada valor-p depende do teste estatístico adotado. Para o teste t de Welch o peso do valor para um nível específico é definido como $[1/N_a + 1/N_b]^{-1}$, em que N_a e N_b, são o total de células nos grupos a e b, respectivamente, no nível especifico do fator. Esse peso é inversamente proporcional à variância esperada do log-*fold change*, assumindo que os grupos e níveis do fator tem a mesma variância. Por outro lado, para a estatística WMW o peso é definido como $N_a\, N_b$, isso implica que os valores-p de níveis com mais células terão uma contribuição maior para o valor-p combinado de cada gene. A estatística (3.10) tem distribuição Normal padrão e portanto o valor-p combinado é obtido calculando de acordo com a hipótese nula alguma probabilidade acumulada da distribuição Normal padrão (WHITLOCK, 2005). Como as comparações múltiplas são realizadas para cada nível do fator conhecido o efeito das estatísticas calculadas, o log-*fold change* para o teste t de Welch ou a AUC para o teste de WMW, são também combinados ponderando pelo mesmo peso do valor-p entre os níveis do fator (LUN; MCCARTHY; MARIONI, 2016). Essa abordagem está implementada nas funções `pairwiseTTests` e `pairwiseWilcox` do pacote **scran**, especificando o fator conhecido no argumento `block`.

Outra abordagem para realizar as comparações múltiplas incorporando fatores de variação conhecidos é utilizar um modelo linear para modelar a expressão gênica de cada gene em função dos grupos e fatores de variação conhecidos. Essa estratégia é bem difundida na literatura da Bioinformática devido principalmente ao pacote **limma** (RITCHIE et al., 2015). Sem perdas de generalidade, supondo apenas um fator categórico de variação conhecido, o modelo linear proposto por Lun, McCarthy e Marioni (2016) para

realizar as comparações múltiplas é definido por

$$y_{ijk} = \alpha_j + \tau_k + \varepsilon_{ijk}, \tag{3.11}$$

com $i = 1, \ldots, n$ células, $j = 1, \ldots, g$ grupos e $k = 1, \ldots, l$ níveis do fator de variação conhecido. Assim, y_{ijk} representa a expressão gênica log-normalizada da célula i, no grupo j do nível k do fator de variação conhecido, α_j é o efeito do grupo j, τ_k é o efeito do nível k do fator de variação conhecido e ε_{ijk} é a componente aleatória do modelo cujo distribuição é Normal com média zero e variância constantes. Importante notar que em sua formulação o modelo não contém um média geral, isto é, o intercepto. Naturalmente, o modelo (3.11) pode ser escrito na notação tradicional dos modelos lineares (SEBER; LEE, 2003), dada por

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{3.12}$$

em que $\boldsymbol{\beta}$ são os parâmetros do modelo representando os níveis dos grupos e fatores conhecidos, $\mathbf{X}$ é a matriz de delineamento e $\boldsymbol{\varepsilon}$ é o erro aleatório de distribuição Normal com média zero e variância constante σ^2.

Para exemplificar, o uso do modelo linear na detecção de genes marcadores, considere um delineamento balanceado com $g = 3$ grupos e $l = 2$ níveis do fator de variação conhecido. Então, a matriz do modelo é dada por

$$\mathbf{X} = \begin{bmatrix} g=1 & g=2 & g=3 & l=1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}. \tag{3.13}$$

No modelo linear, as comparações múltiplas entre os grupos são realizadas utilizando a teoria de contrastes. Um contraste é uma combinação dos coeficientes estimados

do modelo $\boldsymbol{c}^\top\boldsymbol{\beta}$, em que $\boldsymbol{c}$ é um vetor coluna com número de linhas iguais o número de coeficientes do modelo linear. Assim, o objetivo é testar $\mathcal{H}_0 : \boldsymbol{c}^\top\boldsymbol{\beta} = \tau$ ou $\mathcal{H}_0 : \boldsymbol{c}^\top\boldsymbol{\beta} \leqslant (\geqslant)\tau$, em que τ é um valor arbitrário do log-*fold change*. Assumindo que $\tau = 0$, a estatística do teste é dada por:

$$t = \frac{\boldsymbol{c}^\top\widehat{\boldsymbol{\beta}}}{\sqrt{\widehat{\text{Var}}\left[c^\top\widehat{\boldsymbol{\beta}}\right]}} \tag{3.14}$$

em que $\widehat{\boldsymbol{\beta}}$ é a estimativa de mínimos quadrados dos parâmetros do modelo linear, além disso, a variância do denominador pode ser escrita na forma

$$\widehat{\text{Var}}\left[\boldsymbol{c}^\top\widehat{\boldsymbol{\beta}}\right] = \widehat{\sigma}^2\left[\boldsymbol{c}^\top\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\boldsymbol{c}\right]$$

sendo que $\widehat{\sigma}^2$ a variância dos resíduos estimada do modelo linear. Sob a hipótese nula a estatística (3.14) tem distribuição t de Student com $n - p$ graus de liberdade, em que n é o tamanho amostral e p a dimensão do vetor $\boldsymbol{\beta}$. A Tabela 3.1 apresenta os vetores de contrastes $\boldsymbol{c}$ que devem ser utilizados para realizar as comparações múltiplas do exemplo mencionado.

Tabela 3.1 – Vetores de contraste $\boldsymbol{c}$ para as comparações múltiplas entre os $g = 3$ grupos com fator de variação conhecido de $l = 2$ níveis.

Níveis	Contrastes		
	$1-2$	$1-3$	$2-3$
$g = 1$	1	1	0
$g = 2$	-1	0	1
$g = 3$	0	-1	-1
$l = 1$	0	0	0

Sob o ponto de vista estatístico, o modelo linear é mais parcimonioso do que realizar as comparações múltiplas dentro de cada nível dos fatores conhecidos, além disso, permite também a inclusão de covariáveis contínuas, o que é inviável na abordagem anterior. Em contrapartida, se os fatores conhecidos não são consistentes entre os grupos, a variância do resíduo é inflacionada e as estimativas do log-*fold change* serão distorcidas. Além disso, o modelo linear assume que a variância é igual entre os grupos para cada gene, o que nem sempre é uma suposição adequada. Em particular, a presença de grupos em que o gene possui baixa expressão reduzirá a variância residual do modelo para zero, evitando que o modelo penalize gene com alta variância em outros grupos. A abordagem do modelo linear está implementada na função `pairwiseTTests` do pacote **scran**, especificando a matriz do modelo pelo argumento `design`.

3.2.5 Consolidação dos resultados

As abordagens descritas envolvem uma série de comparações entre todos os pares de grupos para identificar de forma abrangente os genes que definem cada grupo. Para todas as comparações múltiplas envolvendo um único grupo, o interesse é consolidar os resultados em um lista de candidatos a genes marcadores. Essa tarefa se torna computacionalmente intensiva de acordo com o número de grupos e genes. Por exemplo, num cenário com $k = 13$ grupos e $g = 2.000$ genes realiza-se um total de $2000 \cdot 13 \cdot 12 \div 2 = 156.000$ testes de comparações múltiplas, os quais fornecem medidas de efeito (log-*fold change* ou AUC) e significância (valor-p) das comparações realizadas. Portanto, inspecionar o resultado de cada comparação para cada gene é inviável.

Nesse sentido, Lun, McCarthy e Marioni (2016) propõem algumas estratégias para consolidação dos resultados que estão disponíveis na função `combineMarkers` do pacote **scran**. A abordagem de caracterização conjunta se concentra em combinações de genes que juntos conduzem a separação de um grupo em relação aos demais. Assim, para cada grupo, os genes são ranqueados dentro de cada comparação múltipla conforme os valores-p. Desse modo, top T genes fornecem um conjunto de marcadores contendo os principais T genes (classificados por significância) em cada comparação múltipla, garantindo a inclusão de genes que podem distinguir entre quaisquer dois grupos. Para exemplificação, considere para um determinado grupo o conjunto de marcadores com $T = 1$. Assim, o conjunto de genes cujo o ranque é menor ou igual a 1 contém os principais genes das comparações múltiplas entre todos os outros grupos. Obviamente, vários genes podem ter o mesmo ranque, pois genes diferentes podem ter a mesma classificação em diferentes comparações múltiplas.

Por outro lado, a abordagem de caracterização única é mais rigorosa, pois considera apenas genes que são expressos diferencialmente em todas as comparações múltiplas envolvendo o grupo de interesse. Para isso, para cada grupo e gene os valores-p das comparações múltiplas são combinados utilizando a técnica de união de interseção de Berger (IUT) proposta por Berger e Hsu (1996). A hipótese nula é que o gene não possui expressão diferente em alguma das comparações, portanto, a hipótese alternativa é que o gene é expresso diferentemente em todos os demais grupos. Na sequência, os genes são ranqueados baseados nos valores-p combinados. Essa estratégia é particularmente eficaz ao lidar com grupos distintos que possuem um perfil de expressão exclusivo. Nesses casos, produz um conjunto de marcadores altamente focado que captura concisamente as diferenças entre os grupos.

É importante ressaltar que os valores-p calculados em cada comparação múltipla para cada gene não podem ser interpretados como uma medida de significância estatística, uma vez que, os grupos encontrados foram identificados empiricamente utilizando o mesmo conjunto de dados. Como os testes usuais não levam em conta a incerteza presente nos

grupos devido ao conjunto de dados, os valores-p são imprecisos. Dessa forma, os valores-p devem ser utilizados apenas para ranquear candidatos a genes marcadores, os quais devem ser validados em estudos posteriores (LUN; MCCARTHY; MARIONI, 2016). Todavia, deve se enfatizar que a avaliação da significância estatística não é uma preocupação em outras análises onde os grupos são pré estabelecidos. Nestes casos, correções do valor-p que levam em consideração o *false dicovery rate* podem ser utilizadas diretamente para definir genes estatisticamente significativos em cada comparação.

3.3 Anotação dos grupos

Seguramente, a tarefa mais desafiadora da análise de dados de scRNA-seq é a interpretação dos resultados. Obter os agrupamentos das células é de certo modo simples e vários métodos estatísticos estão consolidados na literatura para realizar essa tarefa. Contudo, a complexidade é determinar qual estado biológico representa cada agrupamento das células. Para realizar isso, é preciso compreender o conjunto de dados em análise e ter o conhecimento biológico anterior, o que nem sempre é possível de maneira consistente e quantitativa. Por isso, a interpretação da análise é frequentemente conduzida de forma manual com interações de pesquisadores de diferentes áreas (AMEZQUITA et al., 2020).

Para auxiliar nesta etapa da análise existem abordagens computacionais que exploram informações anteriores para atribuir significado às análises realizadas. Em particular, uma abordagem para anotação dos tipos de células é comparar diretamente os perfis de expressão com conjuntos de dados de referência publicados, onde cada célula já foi anotada com seu estado biológico putativo por especialistas no domínio.

Atualmente, vários pacotes estatísticos estão disponíveis para realizar a anotação dos tipos de células. No *software* R, o pacote **Seurat** utiliza um método baseado no vizinho mais próximo para obter as anotações do tipo de célula (BUTLER et al., 2018) a partir dos dados de referência. Os pacotes **SingleR** e **scmap** fornecem funcionalidades dentro da estrutura do Bioconductor para anotar tipos de células usando a correlação entre as expressões de referência e dos dados observados (KISELEV; YIU; HEMBERG, 2018; ARAN et al., 2019). É possível utilizar vários tipos de correlação, entretanto os autores recomendam o uso da correlação de Spearman (ARAN et al., 2019). O pacote **scPred** também utiliza a estrutura do Bioconductor, porém, aplica um modelo de *Support Vector Machine* (SVM) utilizando as componentes principais da expressão gênica para classificar os tipos de células (ALQUICIRA-HERNANDEZ et al., 2019).

Outro pacote do *software* R que tem sido utilizado para anotação do tipo de células predominantes em cada agrupamento é o **clustifyr** (FU et al., 2019). O **clustifyr** suporta tanto a infraestrutura do **Seurat** quanto do Bioconductor e para obter as anotações de cada grupo utiliza a correlação entre as expressões gênicas do conjunto de referência

(com os tipos de células conhecidos) e a matriz da análise. Fu et al. (2019) compararam diferentes correlações e concluíram que a correlação de Spearman fornece melhor resultado.

Importante mencionar que, embora a maioria dos métodos utilizem a correlação de Spearman como medida de similaridade, os métodos possuem particularidades que diferenciam entre si. Uma extensiva comparação entre os métodos citados é realizada por Fu et al. (2019).

3.4 Análise da expressão diferencial gênica

A análise de expressão diferencial (DE) é tarefa fundamental para entendimento das respostas moleculares que ocorrem durante e/ou depois de alguma pertubação ou em diferentes condições de saúde. Desde o surgimento da tecnologia de *microarray* (SCHENA et al., 1995), muitos esforços têm sido posto na proposta de diferentes metodologias estatísticas para análise DE. Uma análise clássica de DE consiste em identificar e mensurar as diferenças nos níveis de expressão gênica de organismos sob diferentes condições experimentais ou fenótipos. No contexto dos dados de *microarray*, abordagens utilizando modelos lineares, em particular, o modelo limma (SMYTH, 2004; RITCHIE et al., 2015; SMYTH, 2005a) tem sido a principal metodologia estatística utilizada. É importante mencionar que para dados de experimentos de *microarray* a abundância de determinado gene é medida por uma intensidade de florescência, o que é efetivamente uma variável contínua, enquanto que para dados dados oriundos de experimentos de *bulk* RNA-seq a abundância é observada como contagem. Portanto, os modelos bem estabelecidos para *microarray*, por exemplo, o limma, não necessariamente são adequados para dados de RNA-seq.

Isso motivou o desenvolvimento de novas estratégia, por exemplo, o modelo limma-voom introduzido por Law et al. (2014) e Liu et al. (2015), uma extensão do modelo limma para dados de RNA-seq, onde os autores propõem pesos específicos para cada amostra com intuito de capturar a variabilidade inerente dos dados de RNA-seq. Outros autores propõem modelar diretamente as contagens observadas assumindo que o processo gerador dos dados pode ser expresso por uma distribuição discreta, usualmente a Binomial Negativa. Nesta classe, tem-se o modelo edgeR desenvolvido nos trabalhos de Robinson e Smyth (2007a), Robinson e Smyth (2007b) e McCarthy, Chen e Smyth (2012) com as implementações computacionais disponíveis no pacote **edgeR** do projeto Bioconductor (ROBINSON; MCCARTHY; SMYTH, 2009). Outra abordagem que também utiliza a distribuição Binomial Negativa é o modelo DESeq2 proposto por Love, Huber e Anders (2014). As duas principais diferenças entre os modelos edgeR e DESeq2 são o fator de normalização utilizado como *offset* dos modelos e na abordagem pra estimação do parâmetro de dispersão da Binomial Negativa utilizando informação de todos os genes. Uma comparação entre os principais modelos, incluindo edgeR e DESeq2, para análise DE

em dados de bulk RNA-seq é realizada por Schurch et al. (2016).

Por outro lado, ao contrário dos dados provenientes de *bulk* RNA-seq, os níveis de expressão observados em experimentos de scRNA-seq apresentam maior ruído e outras particularidades, como presença significativa de contagens iguais a zero, distribuições multimodais e menor tamanho da biblioteca (KORTHAUER et al., 2016). Por causa da baixa quantidade de RNA presente em uma única célula, alguns genes podem não ser detectados, embora sejam expressos, esse fenômeno denominado de efeito *dropouts* auxilia na inflação de zeros. Tendo em vista as características particulares a análise DE em experimentos de scRNA-seq apresenta novos desafios na modelagem estatística da expressão gênica e diversos grupos de pesquisadores tem dedicado esforços para propor métodos apropriados para análise DE em dados de scRNA-seq. Nesse contexto, Kharchenko, Silberstein e Scadden (2014) propuseram a abordagem SCDE que utiliza uma mistura entre Poisson para modelar o evento *dropout* e Binomial Negativa para o nível da expressão, as inferências acerca são conduzidas sob o paradigma Bayesiano a partir da distribuição posteriori dos genes. Já Vallejos, Marioni e Richardson (2015) e Vallejos, Richardson e Marioni (2016) introduzem o modelo BASiCS (*Bayesian analysis of single-cell sequencing*), que também sob o paradigma Bayesiano assume uma estrutura hierárquica Poisson-Gamma para modelar os níveis expressão gênica levando em consideração efeitos técnicos dos dados de scRNA-seq. Finak et al. (2015) propõem o modelo MAST, onde assume que a expressão gênica pode ser modelada por um modelo *hurdle* Normal, sendo que a parte contínua é modelada utilizando a expressões gênicas log-normalizadas. A estimação deste modelo também é realizada sob o paradigma Bayesiano. Trapnell et al. (2014) introduziram a ferramenta Monocle que, na sua versão 3, permite o usuário escolher entre os modelos Binomial Negativo, Poisson ou quasi-Poisson para realizar a análise DE. Schurch et al. (2016), utilizando o paradigma Bayesiano, propõem conduzir a análise DE considerando um modelo de mistura de distribuições Normais baseado no processo de Dirichlet, este modelo é de certa forma bem flexível pois permite detectar mudanças na distribuição da expressão gênica, todavia, o custo computacional é demasiado.

Recentemente, Soneson e Robinson (2018) conduziram um estudo comparando 36 abordagens para análise DE em dados de scRNA-seq. O estudo avaliou empiricamente dados já analisados da literatura e também via simulação. Segundo a comparação realizada pelos autores os métodos desenvolvidos para *bulk* RNA-seq geralmente não apresentaram um desempenho inferior aos métodos específicos para scRNA-seq. Outro estudo mais recente conduzido por Wang et al. (2019) comparou 11 métodos para análise DE, e também concluíram que "os métodos desenvolvidos para dados de scRNA-seq não tendem a mostrar um melhor desempenho em comparação com os métodos projetados para *bulk* RNA-seq".

A maioria dos modelos para análise da expressão diferencial foram desenvolvi-

dos para avaliar diferenças entre conjunto de células provenientes da mesma população, porém sob diferentes condições experimentais ou podem ser utilizados para comparar subpopulações de células (por exemplo, para detecção de genes marcadores, como discutido na Seção 3.2). Nestes modelos, as células são as unidades experimentais e portanto representam a população na qual inferências estatísticas são realizadas. Com o avanço das tecnologias de alto rendimento de scRNA-seq, as medições na expressão gênica pode ser realizada em centenas a milhares de células por amostra (e.g., diferentes indivíduos) sob distintas condições experimentais (e.g., tratamento ou doença). Neste tipo de desenho experimental além da variabilidade dos subconjuntos de células (subpopulações) há também a variabilidade intrínseca à origem na qual a célula pertence (amostra). Ignorar o efeito da amostra na análise de dados pode levar a inferências equivocadas na detecção de genes diferencialmente expressos bem como na intensidade das diferenças estimadas (LUN; MARIONI, 2017; CROWELL et al., 2020).

Uma abordagem que tem sido utilizada para analisar dados com estrutura de dependência entre as células é consolidar as contagens de cada célula em amostras pseudo-*bulk*, em geral, somando as contagens das células no mesmo nível amostral dentro de cada subpopulação das células. Em seguida métodos usuais da análise DE para *bulk* RNA-seq são empregados. Essa abordagem foi realizada por Lun e Marioni (2017) para remover efeitos de células sequenciadas na mesma bandeja (*plates*). Embora pareça contraditório utilizar scRNA-seq para obter amostras pseudo-*bulk*, isso aproveita a resolução oferecida pelas tecnologias de célula única para definir as populações de células e combina com o rigor estatístico dos métodos existentes para análises DE para *bulk* RNA-seq envolvendo um pequeno número de amostras. Recentemente, Crowell et al. (2020) desenvolvoram a ferramenta **muscat** que contém várias abordagens, incluindo a agregação em amostras pseudo-*bulk*. Além disso, o trabalho de Crowell et al. (2020) também apresentam modelos de efeito mistos para capturar a estrutura de variância presente nos dados.

3.4.1 Modelo edgeR

No contexto dos experimentos de RNA-seq, seguramente o modelo Binomial Negativo proposto por McCarthy, Chen e Smyth (2012) e disponível no pacote **edgeR** é uma das principais abordagens na literatura para análise DE. Dado um particular gene g, considere que y_{gi} denota a contagem da i-ésima amostra. Importante ressaltar que nos dados provenientes de scRNA-seq com estrutura de dependência dentro das células (amostras) as contagens observadas são somadas nos níveis apropriados resultando em amostras denominadas de pseudo-*bulk*. Os autores assumem que as contagens observadas são descritas por uma modelo linear generalizado (GLM) com distribuição Binomial Negativa, ou seja,

$$y_{gi} \sim \text{BN}(\mu_{gi}, \phi_g) \tag{3.15}$$

em que $\mu_{gi} = N_i\,\lambda_{gi}$ é a média e $\phi_g > 0$ o parâmetro de dispersão, sendo N_i o tamanho da biblioteca (soma das contagens) na amostra i e λ_{gi} a abundância relativa do gene g na amostra i, ou seja, a proporção esperada de contagens ("*reads*") mapeadas no gene g. Assumindo que λ_{gi} é relacionado aos preditores por meio da relação log-linear. Assim, a esperança de y_{gi} é dada por:

$$\log \mu_{gi} = \log \lambda_{gi} + \log N_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}_g + \log N_i$$

em que $\boldsymbol{x}_i = (x_{1i}, \ldots, x_{pi})$ é o vetor de preditoras, $\boldsymbol{\beta}_g = (\beta_{g1}, \ldots, \beta_{gp})$ é o vetor de coeficientes representam o efeito das preditoras. Nesta parametrização a variância de y_{gi} é dada por $\mu_{gi} + \mu_{gi}^2\phi_g)$, portanto o modelo Binomial Negativo acomoda casos de super-dispersão quando $\phi_g > 0$, se $\phi_g = 0$ o modelo reduz ao modelo de Poisson.

Implicitamente, o modelo GLM definido em (3.15) assume que o parâmetro de dispersão ϕ_g é conhecido ou estimado a priori. Assim, a estimação do vetor de parâmetros $\boldsymbol{\beta}_g$ é realizada pelo método escore de Fisher (NELDER; WEDDERBURN, 1972). Especificamente, o edgeR aprimora o algoritmo escore de Fisher com uma modificação de amortecimento de Levenberg disponível no pacote **statmod** (SMYTH, 2005b). O algoritmo modificado força uma redução na função *deviance* a cada iteração. Assim, a sequências de *deviances* se torna monótona e limitada, portanto, sempre converge, a menos que imprecisões de ponto flutuante intervenham primeiro (CHEN; LUN; SMYTH, 2014).

A estimativa precisa do parâmetro de dispersão ϕ_g no modelo Binomial Negativo é fundamental para ajustar os GLMs e avaliar a expressão diferencial em cada gene. Em particular, McCarthy, Chen e Smyth (2012) mostram que o parâmetro $\phi_g^{1/2}$ representa o coeficiente de variação biológico, isto é, a variabilidade com a qual a verdadeira abundância do gene varia entre as amostras de RNA. Em outras palavras, ϕ_g representa a variabilidade se a precisão do sequenciamento pudesse ser aumentada indefinidamente. Assim, segundo os autores, ϕ_g representa a principal fonte de variação para contagens altas dos genes, e portanto estimativas confiáveis para ϕ_g são primordiais. A estimativa de máxima verossimilhança para ϕ_g apresenta um desempenho ruim em amostras pequenas, o que geralmente é o caso dos dados de RNA-seq ou as amostras pseudo-*bulk* obtidas (ROBINSON; SMYTH, 2007b). Em particular, segundo Robinson e Smyth (2007b) as estimativas de máxima verossimilhança tendem a subestimar o parâmetro de dispersão.

Para experimentos com apenas um fator Robinson e Smyth (2007a) e Robinson e Smyth (2007b) discutem o método de máxima verossimilhança condicional com ajuste de quantis para a estimação do parâmetro de dispersão. A estimação no caso geral é estudada por McCarthy, Chen e Smyth (2012) que propõem o método baseado na verossimilhança perfilada proposta por Cox e Reid (1987). A condição do método de Cox e Reid (1987) é que o parâmetro de incômodo (*nuance*) seja ortogonal ao parâmetro de interesse. Na estimação do parâmetro de dispersão ϕ_g é o parâmetro de interesse e $\boldsymbol{\beta}_g$ é o parâmetro de incômodo e a condição de ortogonalidade é satisfeita (SMYTH; VERBYLA, 1999).

A função log-verossimilhança perfilada ajustada pelo método de Cox e Reid (1987) (APL) para o parâmetro ϕ_g é definida por

$$\text{APL}_g(\phi_g) = \ell(\phi_g \mid \boldsymbol{y}_g, \widehat{\boldsymbol{\beta}}_g) - \frac{1}{2}\log\det(\mathcal{I}_g) \tag{3.16}$$

em que $\ell(\phi_g \mid \boldsymbol{y}_g, \widehat{\boldsymbol{\beta}}_g)$ é a função log-verossimilhança do modelo Binomial Negativo avaliada nos dados observados $\boldsymbol{y}_g$ e na estimativa $\widehat{\boldsymbol{\beta}}_g$ do vetor de parâmetros, $\mathcal{I}_g$ é a informação de Fisher do modelo Binomial Negativo. Note que $\widehat{\boldsymbol{\beta}}_g$ e $\mathcal{I}_g$ são funções de ϕ_g, assim a maximização da função (3.16) fornece uma estimativa para ϕ_g.

Uma técnica que tem sido bastante utilizada na análise DE em dados de genoma é o Bayes empírico (*empirical Bayes*). Brevemente, está técnica tem como objetivo estimar a distribuição a priori dos dados e utilizá-la na abordagem Bayesiana para obter a distribuição a posteriori. Estudos mostram que a estimação utilizando o método Bayes empírico tem fornecido melhores estimativas do que o método da verossimilhança para dados em alta dimensão (SMYTH, 2004; EFRON et al., 2001; CHEN; LUN; SMYTH, 2014). Não há distribuição a priori conjugada para o parâmetro de dispersão ϕ_g da Binomial Negativa e a aplicação direta do Bayes empírico não é possível. Para contornar este problema, os autores propuseram verossimilhanças ponderadas, uma vez que a estimativa obtida pelo método Bayes empírico é equivalente à estimativa obtida maximizando uma verossimilhança ponderada (WANG, 2006). Chen, Lun e Smyth (2014) discutem a teoria de três abordagens disponíveis no pacote **edgeR** para a estimação do parâmetro de dispersão utilizando versões ponderadas da verossimilhança definida em (3.16).

A abordagem mais simples para compartilhar informações entre os genes é assumir que todos os genes têm o mesmo parâmetro de dispersão ϕ. A estimação é então realizada maximizando a log-verossimilhança APL comum, definida por:

$$\text{APL}_C(\phi) = \frac{1}{G}\sum_{i=1}^{G}\text{APL}_g(\phi) \tag{3.17}$$

em que G é total de genes. Note que a função $\text{APL}_C(\phi)$ atribui pesos iguais para as verossimilhanças de todos os genes.

A suposição de dispersão comum não retrata a natureza dos dados. De fato, em muitos conjuntos de dados de RNA-seq, genes com baixo nível de expressão tendem a ter alta dispersão e vice versa (CHEN; LUN; SMYTH, 2014). Assim, uma suposição razoável é assumir que os valores de dispersão dependem do nível de expressão do gene e essa relação pode ser expressa por um modelo de tendência média-dispersão. Essa abordagem discutida em detalhes por Chen, Lun e Smyth (2014) no contexto de RNA-seq foi proposta inicialmente por (ROBINSON; SMYTH, 2007b) para analisar dados de *digital gene expression* (DGE).

Conforme descrevem Chen, Lun e Smyth (2014) o nível de expressão geral para cada gene é estimado como a média entre as amostras e expresso na escala logaritmo da

contagem média por milhão (logCPM), em que a contagem média é estimada utilizando um modelo Binomial Negativo definido em (3.15), porém com dispersão comum. Os genes são ordenados de acordo com a média dos valores de logCPM. Para um particular gene g, uma função APL local global denotada por $\text{APL}_{S_g}(\phi_g)$ é construída calculando a média das função APLs de um conjunto de genes C_g. Segundo Chen, Lun e Smyth (2014) o conjunto C_g contém pelo menos 25% dos genes que são mais próximos conforme a média do logCPM do gene g.

O peso de cada gene no conjunto C_g é especificado para levar em consideração a relevância no nível de expressão do gene g entre os outros genes do conjunto C_g. Em particular, a contribuição (peso) do gene a pertencente ao conjunto C_g na função APL é determinado por

$$\omega_a = (1 - |x_a|^3)^3,$$

em que $-1 < x_a < 1$ denota a diferença média padronizada no intervalo $(-1, 1)$ do logCPM entre os genes g e a. Assim, quanto mais próximo o nível de expressão entre os genes g e a menor $|x_a|$ e portanto maior o peso ω_a. Dessa forma, a estimativa de ϕ_g modelando a relação média-dispersão é obtida maximizando a função log-verossimilhança $\text{APL}_{S_g}(\phi_g)$ é definida por

$$\text{APL}_{S_g}(\phi_g) = \frac{\sum\limits_{a \in C_g} \omega_a \cdot \text{APL}_a(\phi_g)}{\sum\limits_{a \in C_g} \omega_a}. \tag{3.18}$$

A estimação do parâmetro de dispersão modelando a relação média-dispersão não é suficiente na prática uma vez que a dispersão é especifica do comportamento do gene, e portanto uma estimativa individual para a dispersão para cada gene deve ser realizada. Para obter estimativas do parâmetro de dispersão estáveis para cada gene Robinson e Smyth (2007a) propõem uma versão da verossimilhança baseado no princípio do Bayes empírico que combina informações individuais de cada gene e compartilhadas pela relação média-variância definida em (3.18). Tal abordagem tem o efeito de espremer as dispersões de cada gene em direção a uma estimativa combinada, resultando em inferências mais estável quando o número de amostras é pequeno (CHEN; LUN; SMYTH, 2014). A versão ponderada da verossimilhança APL para estimar a dispersão especifica do gene g, é definida por

$$\text{APL}_{W_g}(\phi_g) = \text{APL}_g(\phi_g) + G_0 \cdot \text{APL}_{S_g}(\phi_g) \tag{3.19}$$

em que $\text{APL}_g(\phi_g)$ é a contribuição utilizando a informação apenas do gene g, $\text{APL}_{S_g}(\phi_g)$ é a contribuição utilizando informações da relação média-dispersão e G_0 é o peso atribuído a $\text{APL}_{S_g}(\phi_g)$.

A escolha ótima de G_0 vai depender da variabilidade das dispersões, sendo que valores altos melhores quando a dispersão é constante para todos os genes ou segue a tendencia da relação média-dispersão. Quando a dispersão varia bastante entre os

genes, valores pequenos são recomendáveis. O pacote **edgeR** utiliza abordagem de quase-verossimilhança proposta por Robinson e Smyth (2007b) para estimar de forma robusta o valor de G_0. Detalhes são discutidos por Chen, Lun e Smyth (2014).

Após a estimação dos parâmetros do modelo edgeR testes de hipóteses são conduzidos para avaliar a hipótese de diferença no nível de expressão entre amostras sob diferentes condições experimentais ou fenótipos. De forma geral, a hipótese nula pode ser formulado da seguinte forma

$$\mathcal{H}_0 : \boldsymbol{c}^\top \boldsymbol{\beta}_g = 0 \tag{3.20}$$

em que $\boldsymbol{c}^\top$ é um vetor de contrastes.

McCarthy, Chen e Smyth (2012) propõem utilizar o teste da razão de verossimilhanças para testar a hipótese (3.20). A estatística da razão de verossimilhanças é dada por

$$S_{LR} = 2\left[\ell\left(\widehat{\boldsymbol{\beta}}_g \mid \boldsymbol{y}_g, \widehat{\phi}_g\right) - \ell\left(\widetilde{\boldsymbol{\beta}}_g \mid \boldsymbol{y}_g, \widehat{\phi}_g\right)\right] \tag{3.21}$$

em que $\widehat{\boldsymbol{\beta}}_g$ e $\widetilde{\boldsymbol{\beta}}_g$ são as estimativas de máxima verossimilhança obtida sob a hipótese alternativa e nula, respectivamente, para o vetor de parâmetros $\boldsymbol{\beta}$. Sob a hipótese nula a estatística S_{LR} tem distribuição assintótica qui-quadrado com p graus de liberdade, em que p é a dimensão do parâmetro $\boldsymbol{\beta}$.

3.4.2 Modelo linear misto

Uma alternativa para modelar diretamente as expressões das células na análise DE em dados de scRNA-seq quando há estrutura de dependência entre as observações, em particular, células provenientes de diferentes origens, é utilizar a classe de modelos mistos. Lun e Marioni (2017) utilizaram o modelo Binomial Negativo misto para modelar diretamente as contagens incorporando o efeito de bandeja presente no experimento. No contexto de RNA-seq Hoffman e Roussos (2020) utilizaram modelos mistos para modelar as expressões $\log_2$-normalizadas e ponderadas em delineamentos com medidas repetidas. Recentemente, Crowell et al. (2020) disponibilizaram no Bioconductor o pacote **muscat** que fornece três abordagens para utilização de modelos mistos: (i) ajustar modelos lineares mistos nas expressões $\log_2$-normalizadas com pesos nas observações, por exemplo, utilizar os pesos do método voom (LAW et al., 2014); (ii) ajustar modelos lineares mistos utilizando as expressões normalizadas por métodos de estabilização da variância, por exemplo, a proposta de Hafemeister e Satija (2019); e (iii) ajustar modelos generalizados mistos diretamente nas contagens considerando as distribuições Poisson ou Binomial Negativa.

Considere um experimento de scRNA-seq onde existem n indivíduos e para cada um a expressão gênica de $g = 1, \ldots, r$ genes é mensurada em m_i células, com $i = 1, \ldots, n$. Seja $\boldsymbol{y}_{gi} = (y_{gi1}, \ldots, y_{gim_i})$ as expressões do gene g $\log_2$-normalizadas do i-ésimo indivíduo.

O modelo linear misto é definido por

$$\boldsymbol{y}_{gi} = \mathbf{X}_i\boldsymbol{\beta}_g + \mathbf{Z}_i\mathbf{b}_{gi} + \boldsymbol{\varepsilon}_{gi}, \quad i = 1, \ldots, n \tag{3.22}$$

en que $\boldsymbol{\beta}_g$ é um vetor de dimensão $p \times 1$ de parâmetros dos efeitos fixos para o gene g, $\mathbf{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{ip})$ é uma matriz de especificação dos efeitos fixos com dimensão $m_i \times p$, conhecida e de posto completo, em que $\boldsymbol{x}_{ij} = (x_{ij1}, \ldots, x_{ijm_i})^\top$ representa o vetor com os m_i valores da j-ésima variável independente, $j = 1, \ldots, p$ para o i-ésimo individuo, $\mathbf{b}_{gi}$ é um vetor com dimensão $q \times 1$ dos efeitos aleatórios, os quais refletem o comportamento particular do i-ésimo indivíduo na expressão do gene g, $\mathbf{Z}_i$ é uma matriz conhecida de especificação dos efeitos aleatórios com dimensão $m_i \times q$ e posto completo e $\boldsymbol{\varepsilon}_{gi}$ é um vetor de erros aleatórios do gene g com dimensão $m_i \times 1$.

Supõe-se também que $\mathbf{b}_{gi} \sim N(\mathbf{0}, \mathbf{G})$ e $\boldsymbol{\varepsilon}_{gi} \sim N(\mathbf{0}, \mathbf{R}_{gi})$, em que $\mathbf{G}$ tem dimensão $q \times q$ e $\mathbf{R}_i$ com dimensão $m_i \times m_i$ são matrizes simétricas definidas e positivas, além disso, $\mathbf{b}_{ig}$ e $\boldsymbol{\varepsilon}_{gi}$ são variáveis aleatórias independentes. Sob este modelo, o vetor de expressões gênicas do i-ésimo indivíduo tem distribuição Normal multivariada com média e matriz de covariâncias dadas, respectivamente, por:

$$\mathbb{E}(\boldsymbol{y}_{gi}) = \mathbf{X}_i\,\boldsymbol{\beta}_g \tag{3.23}$$

e

$$\mathrm{Var}(\boldsymbol{y}_{gi}) = \mathbf{V}_{gi} = \mathbf{Z}_i\,\mathbf{G}\,\mathbf{Z}_i^\top + \mathbf{R}_{gi}. \tag{3.24}$$

Na modelagem de dados com medidas repetidas, o grande desafio é especificar a a estrutura da matriz de covariâncias. Usualmente, a especificação da matriz de covariâncias vai depender da forma de coleta dos dados e do processo gerador das expressões gênicas. Diggle et al. (2002) sugerem que a matriz de covariâncias deve ser suficientemente flexível a ponto de incluir no minimo três fontes de variação aleatória: (i) devida a efeitos aleatórios, quando as unidades de medidas formam uma amostras aleatória da população de interesse; (ii) a variação que pode ser explicada por um correlação serial e (iii) a variação devido a erros de medida.

No contexto dos dados de scRNA-seq na presença de células de diferentes indivíduos duas fontes de variação devem ser levadas em consideração. Inicialmente, devido a erros de medidas e também a especificação do modelo, uma vez que a expressão gênica $\log_2$-normalizada apresenta uma distribuição com caudas pesadas e assimétrica. Dessa forma, para contornar este fato Law et al. (2014) argumentam que é mais importante modelar a relação de média-variância corretamente do que especificar a distribuição probabilística exata das expressões gênica e para isso os autores propõem estimar pesos para cada observação de forma robusta e não paramétrica modelando a relação média-variância das contagens e incorporando os pesos na matriz $\mathbf{R}_{gi}$. Dessa forma, assume-se que

$$\mathbf{R}_{gi} = \sigma_g^2\,\mathrm{diag}(\omega_{gi})$$

em que $\sigma_g^2 > 0$ e ω_{gi} são pesos conhecidos do i-ésimo indivíduo para gene g e estimados de forma não paramétrica modelando relação média-variância das contagens conforme a proposta de Law et al. (2014) .

Para especificação da matriz dos efeitos aleatórios devido ao indivíduo a estrutura uniforme fornece uma especificação parcimoniosa e interpretável. Assumindo a estrutura uniforme e considerando $m_i = 4$ células a matriz de covariância entre as células do mesmo indivíduo é dada por

$$\text{Var}(\boldsymbol{y}_{gi}) = \begin{bmatrix} \sigma_g^2\,\omega_{g1} + \tau & \tau & \tau & \tau \\ \tau & \sigma_g^2\,\omega_{g2} + \tau & \tau & \tau \\ \tau & \tau & \sigma_g^2\,\omega_{g3} + \tau & \tau \\ \tau & \tau & \tau & \sigma_g^2\,\omega_{g4} + \tau \end{bmatrix}.$$

em que $\tau > 0$ é a variância devida ao efeito aleatório de indivíduo. Nesta estrutura duas células 1 e 2 do mesmo indivíduo estão correlacionadas positivamente com coeficiente de correlação dado por

$$\rho_{12} = \frac{\tau}{\sqrt{(\sigma_g^2\,\omega_{g1} + \tau)(\sigma_g^2\,\omega_{g2} + \tau)}}.$$

A estimação dos parâmetros de efeitos fixos $\boldsymbol{\beta}_g$, variância dos erros σ_g^2 e a variância do efeito aleatório τ pode ser realizada sob o paradigma frequentista utilizando os métodos da máxima verossimilhança ou máxima verossimilhança restrita. Em geral os estimadores de máxima verossimilhança são não viesados para os parâmetros de efeitos fixos, porém fornecem estimativas viesadas para as variâncias dos efeitos aleatórios (PINHEIRO, 1994; BATES et al., 2015). Com exceção de alguns modelos particulares, geralmente não é possível obter expressões analíticas para os estimadores dos parâmetros do modelo linear misto. Assim, métodos iterativos são utilizados para, no software R o pacote **lme4** fornece uma implementação em `C++` do algoritmo de mínimos quadrados penalizados para estimação, detalhes teóricos são discutidos por Bates et al. (2015).

Na análise de expressão gênica o usual é a matriz de especificação dos efeitos fixos incluir variáveis *dummies* indicando o grupo a qual a célula pertence, assim após a estimação dos parâmetros o interesse reside em testar hipóteses acerca dos parâmetros, em particular verificar se existe diferença nas expressões entre os grupos. Sem perdas de generalidades, podemos formular a hipótese nula da seguinte forma:

$$\mathcal{H}_0 : \boldsymbol{c}^\top \boldsymbol{\beta}_g = \boldsymbol{0} \tag{3.25}$$

em que $\boldsymbol{c}$ é um vetor de contrastes com dimensão $p \times 1$. Para testar a hipótese (3.25) Giesbrecht e Burns (1985) propõem o utilizar a estatística t, definida por

$$t = \frac{\boldsymbol{c}^\top\,\widehat{\boldsymbol{\beta}}_g}{\sqrt{\boldsymbol{c}\,\widehat{\boldsymbol{\Sigma}}_g\,\boldsymbol{c}^\top}} \tag{3.26}$$

em que $\widehat{\boldsymbol{\Sigma}}_g$ é a matriz de covariâncias estimada dos efeitos fixos $\boldsymbol{\beta}_g$. De forma geral, se assumindo que a variância de $\boldsymbol{y}_{gi}$ é definida em (3.24), então a matriz de covariâncias estimada dos efeitos fixos é dada por

$$\widehat{\boldsymbol{\Sigma}}_g = \left[\mathbf{X}^\top \widehat{\mathbf{V}}_g^{-1} \mathbf{X}\right]^{-1}.$$

em que $\mathbf{X} = (\mathbf{X}_1^\top, \ldots, \mathbf{X}_n^\top)^\top$ e $\mathbf{V}_g = (\mathbf{V}_{g1}^\top, \ldots, \mathbf{V}_{gn}^\top)^\top$.

Giesbrecht e Burns (1985) utilizam a aproximação pelo método dos momentos de Satterthwaite para obter graus de liberdade adequados para a estatística (3.26), definido como

$$\nu = \frac{2\left(\boldsymbol{c}^\top \widehat{\boldsymbol{\Sigma}}_g \boldsymbol{c}\right)^2}{\text{Var}\left(\boldsymbol{c}^\top \widehat{\boldsymbol{\Sigma}}_g \boldsymbol{c}\right)}$$

em que a variância do denominador é aproximada utilizando o método delta (KUZNETSOVA; BROCKHOFF; CHRISTENSEN, 2017).

3.4.3 Interpretação do logFC

Na análise de dados dados Omics (Genoma, Transcriptoma e Proteoma) uma medida muito usual para comparar a expressão gênica entre dois grupos é a mudança em log-razão (log-*fold change* – logFC). O logFC é bastante utilizado devido a interpretabilidade que fornece, possibilitando melhor compreendimento por pesquisadores da genética. Suponha que a expressão do gene g no grupo A seja $x_{gA} = 1.5$ e no grupo B seja $x_{gB} = 2.0$. Neste caso, o *fold change* (mudança na razão) do gene g para o grupo A em relação ao grupo B é dado por $\text{FC}(A) = \dfrac{x_{gA}}{x_{gB}} = \dfrac{1.5}{2.0} = 0.75$. Analogamente, o *fold change* do grupo B é dado por $FC(B) = \dfrac{x_{gB}}{x_{gA}} = \dfrac{2.0}{1.5} = 1.33$. Esses valores indicam que o gene g no grupo A é 0.75 menos regulado do que no grupo B, ou seja, a expressão do gene g em A é 75% da expressão observada em B. Em outras palavras, a expressão do gene g no grupo B é 33% maior que no grupo A. Observa-se que os valores 0.75 e 1.33 são menos interpretáveis para indicar alta ou baixa regulação do gene. Além disso, o *fold-change* varia entre 0 e 1 em comparações com baixa regulação, enquanto que varia entre 1 ao infinito em comparações com alta regulação.

Por este motivo, a função logarítmica é usualmente adotada para entendimento do *fold change*, sendo o logaritmo na base 2 o mais comumente utilizado (LOVE; HUBER; ANDERS, 2014), devido a facilidade de interpretação. Por exemplo, uma duplicação na escala original é igual a $\log_2$-*fold change* de 1, uma quadruplicação é igual $\log_2$-*fold change* de 2 e assim por diante. Além disso, a escala $\log_2$ é simétrica quando a mudança diminui ou aumenta em um valor equivalente. Para o exemplo descrito temos que $\log_2 \text{FC}(A) = -0.415$ e $\log_2 \text{FC}(B) = 0.415$.

A Figura 3.1 ilustra a facilidade de interpretação do logFC em relação ao *fold change*, cada ponto representa o logFC ou *fold change* de um suposto gene $(1, 2, \ldots, 50)$. Neste conjunto de dados, metade dos genes foram regulados positivamente e a outra metade regulados negativamente. Interpretar a mudança de dobra não transformada é complicado. Em particular, parece que Gene 1 teve um grande *fold change*, perto de 100, mas não se pode dizer o mesmo para os genes 30 a 50. Por outro lado, observando o gráfico do logFC percebe-se claramente as diferenças na expressão para os genes 30 a 50.

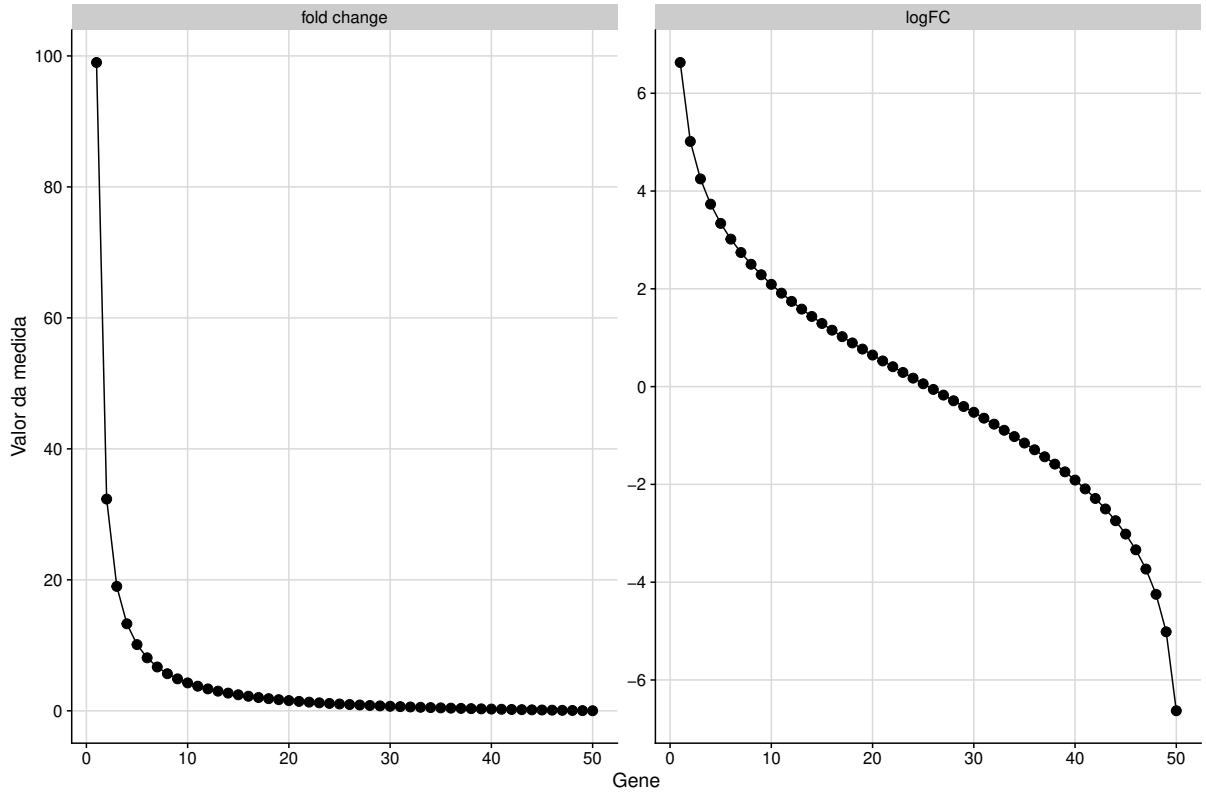


Figura 3.1 – Exemplo fictício do *fold change* e logFC para os genes 1 a 50. Cada ponto representa o *fold change* (à esquerda) e logFC (à direita).

Um representação gráfica bastante utilizada para reportar os resultados das análise é o *volcano plot*, apresentado na Figura 3.2. Neste gráfico de pontos, o eixo x representa o logFC do gene enquanto que o eixo y o é o valor-p na escala $-\log_{10}$ associado ao teste de comparação múltipla, em geral algum teste de médias. Com base nessa representação gráfica podemos identificar genes que são diferentes em termos do valor-p do teste, do logFC ou ambos. Por exemplo, pode-se adotar como de significância limiar se $|\log \mathrm{FC}| \geqslant 2$.

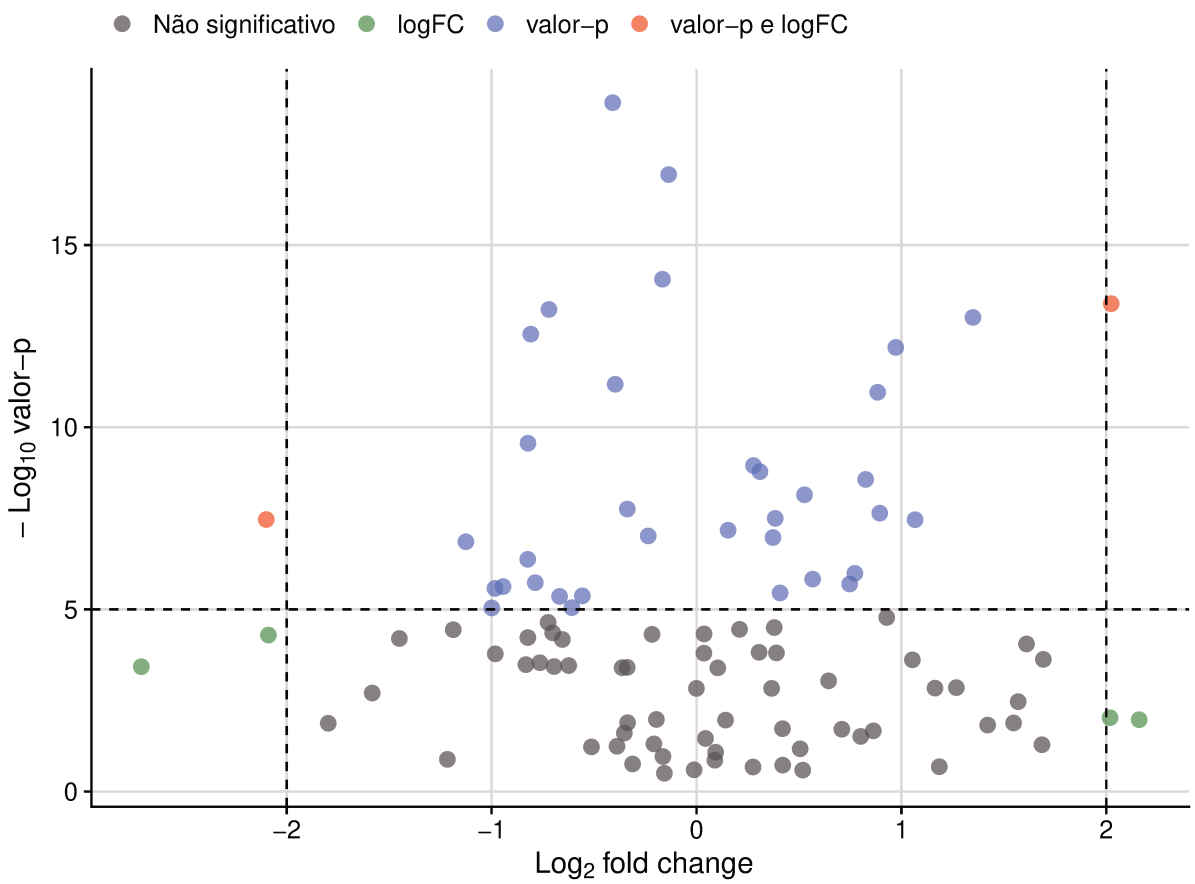


Figura 3.2 – Exemplo fictício do *volcano plot*. Cada ponto representa um gene.

4 Análise de Dados

Como forma de aplicar os conceitos discutidos ao longo da dissertação o presente capítulo apresenta dois estudos. Na Seção 4.1 um estudo de simulação é conduzido para comparar propriedades estatísticas dos modelos discutidos na análise de expressão gênica quando as células sequenciadas são oriundas de diferentes organismos, por exemplo, indivíduos. Além disso, a Seção 4.2 apresenta uma análise de dados reais utilizando o conjunto de dados utilizado por Liao et al. (2020) e cujo os objetivos foram (i) caracterizar e propor novas subpopulações de células e (ii) comparar o nível de expressão gênica entre as condições dos indivíduos.

4.1 Estudo de simulação

Nesta seção discuto de forma geral sobre a estrutura probabilística dos dados de scRNA-seq, em particular, quando a origem da células é diferente. Tomando como exemplo o conjunto de dados de células BALF em indivíduos com COVID-19, uma formulação probabilística análoga à discutida por Lun e Marioni (2017) dos efeitos inerentes aos dados é apresentada na Seção 4.1.1. Na Seção 4.1.2 os parâmetros da formulação proposta são estimados a partir do conjunto de dados reais apresentado por Liao et al. (2020) afim de mimetizar as características particulares dos dados de scRNA-seq e ao mesmo tempo manter o controle sobre o processo gerador dos dados. Com os parâmetros estimados, cenários das simulação são estabelecidos e propriedades teóricas dos modelos são investigadas e discutidas nas Seções 4.1.3 e 4.1.4, respectivamente.

4.1.1 Formulação probabilística da estrutura dos dados

O fluxo usual nas análises de dados de scRNA-seq, após o pré-processamentos das amostras, é descobrir populações de células utilizando técnicas de agrupamento, e na sequência realizar a análise de diferenciação na expressão gênica dentro das populações encontradas. Em grandes estudos, é comum observar dados que embora sejam do mesmo material genético, são oriundos de diferentes organismos. Para contextualizar com a análise de dados reais conduzida na Seção Seção 4.2, considere que os organismos são seres humanos, isto é, o efeito da origem da célula seja devido ao indivíduo. No entanto, o estudo de simulação apresentado aqui pode ser generalizado para qualquer efeito de origem da célula. Por exemplo, o efeito de bandeja discutido por Lun e Marioni (2017). Assumindo que o efeito da origem da célula existe e é significativo, então a expressão de cada gene em todas células de um dado indivíduo é modificada de alguma maneira específica pelos

efeitos do indivíduo e gene. O efeito de gene é motivado pela presença de variabilidade biológica entre as células da mesma população.

Sem perdas de generalidades, supondo apenas uma população de células, considere a variável aleatória Y_{gijk} representando a contagem observada do gene g na célula i do indivíduo j no grupo k. Seja δ_{gjk} uma variável aleatória representando o efeito do gene-indivíduo para o gene g no indivíduo j do grupo k, assumindo que δ_{gjk} para cada indivíduo é independente de uma distribuição com suporte positivo, média um e variância não nula. Seja θ_{ijk} a variável aleatória representando o viés da célula i no indivíduo j do grupo k, também oriunda de uma distribuição positiva com média unitária. Além disso, assume-se que θ_{ijk} e δ_{gjk} são independentes. Podemos supor também que a variável aleatória Y_{gijk} é observada independentemente para cada gene e célula, condicionada nos valores observados de θ_{ijk} e δ_{gjk} com esperança condicional dada por:

$$\mathbb{E}(Y_{gijk} \mid \delta_{gjk}, \theta_{ijk}) = \delta_{gjk}\, \theta_{ijk}\, \mu_{gk}$$

em que μ_{gk} é o valor esperado da contagem do gene g no grupo k. Essa formulação representa o impacto dos indivíduos nos dados.

Um modelo probabilístico que tem sido comum na literatura para simular as contagens de scRNA-seq é a distribuição Binomial Negativa (BN). Embora alguns estudos argumentam que para capturar o efeito *dropout* é necessário a versão inflacionada de zeros, por exemplo, Vallejos, Marioni e Richardson (2015), Finak et al. (2015), Risso D.and Perraudeau et al. (2018), recentemente Svensson (2020) e Townes et al. (2019) mostraram empiricamente que o efeito *dropout* pode ser descrito pela distribuição BN. Nesse sentido, para simular as contagens com efeito de indivíduo, a formulação proposta por Lun e Marioni (2017) foi utilizada, sendo que

$$Y_{gijk} \mid \delta_{gjk}, \theta_{ijk} \sim \text{BN}(\delta_{gjk}\, \theta_{ijk}\, \mu_{gk}, \phi_g). \tag{4.1}$$

Supondo que $\log(\delta_{gjk})$ segue uma distribuição Normal com média $\sigma^2/2$ e variância σ^2 de tal forma que $\mathbb{E}(\delta_{gjk}) = 1$, ou seja, o efeito médio de indivíduo para cada gene atinge média igual a um. O parâmetro θ_{ijk} é amostrado de uma empírica obtida do conjunto com média igual a um. Assume-se que os parâmetros δ_{gjk} e θ_{ijk} tem esperança igual a um por, assim é possível garantir que $E(Y_{gijk}) = \mu_g$, consistente com a definição marginal da distribuição de Y_{gijk}.

Importante ressaltar que outras formulações probabilísticas para simulação de que buscam caracterizar particulares dos dados de scRNA-seq podem ser encontradas na literatura. O pacote **splater** disponível no Bioconductor e introduzido por Zappia, Phipson e Oshlack (2017) apresenta as principais formulações para simulação de dados de scRNA-seq discutidas na literatura, incluindo a proposta de Lun e Marioni (2017).

4.1.2 Estimação dos parâmetros a partir dos dados

Afim de gerar os dados com estrutura mais realística possível, os parâmetros da formulação (4.1) foram estimados considerando as amostras de células dos indivíduos HC1, HC2, M1, M2, S1 e S2. O pré-processamento empregado por Liao et al. (2020) para remoção de células e genes de baixa qualidade e a seleção de genes descrita em detalhes na Seção 4.2.1 foi conduzida antes da estimação dos parâmetros. Dois conjuntos de parâmetros foram estimados, considerando apenas os indivíduos HC1, M1 e S1, e considerando todos os seis indivíduos.

Para estimação do viés da célula θ_{ijk} o método *deconvolution* (LUN; BACH; MARIONI, 2016) foi utilizado para obter os fatores de escala para cada células. A Figura 4.1 apresenta as densidades do fator de escala conforme o indivíduo. Nota-se que para os indivíduos M1 e M2 os fatores de escalas apresentam uma distribuição bimodal. Como mencionado, a média do fator de escala entre todos os indivíduos é igual a um.

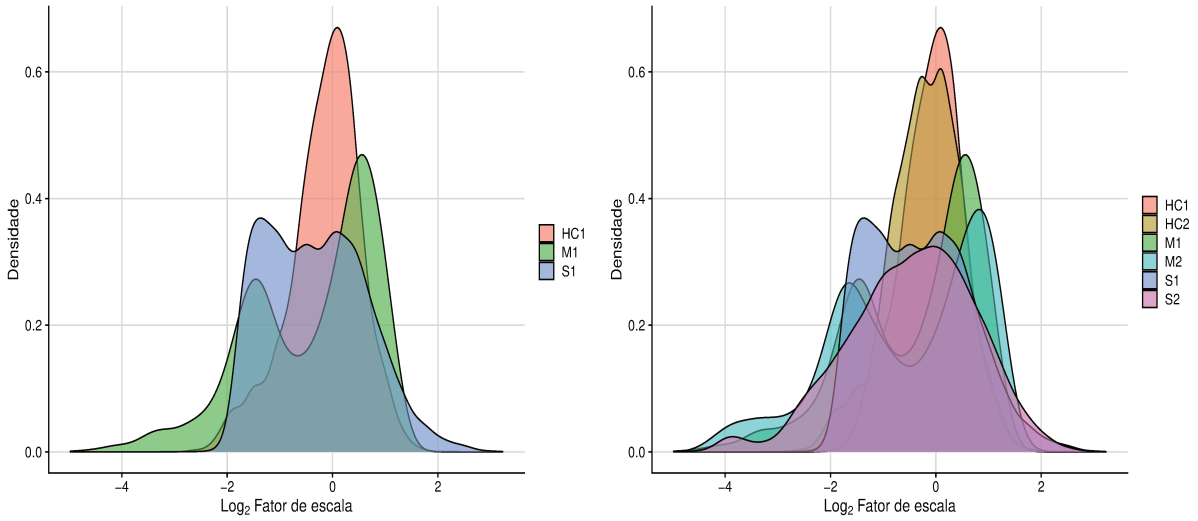


Figura 4.1 – Distribuição do fator de escala conforme o indivíduo. Gráfico esquerda considerando 3 indivíduos e à direita 6 indivíduos.

Para estimar a contagem média de cada gene, um modelo binomial negativo apenas com intercepto foi ajustado utilizando a função `mglmOneGroup` do pacote **edgeR** (ROBINSON; MCCARTHY; SMYTH, 2009) com *offset* centrada no fator de escala. Os coeficientes obtidos foram utilizados para definir a contagem esperada para cada gene μ_g, ver Figura 4.2. Os genes DE foram definidos por $\mu_{gk} = \psi_{gk}\,\mu_g$, para algum *fold change* ψ_{gk}, aqueles que não são DE considera-se $\psi_{gk} = 1$. A estimação do parâmetro de dispersão foi realizada seguindo a proposta de Chen, Lun e Smyth (2014), também disponível no **edgeR** através da função `estimateDisp`. A estimação dos parâmetros de média e dispersão foram realizados condicionando ao efeito do indivíduo.

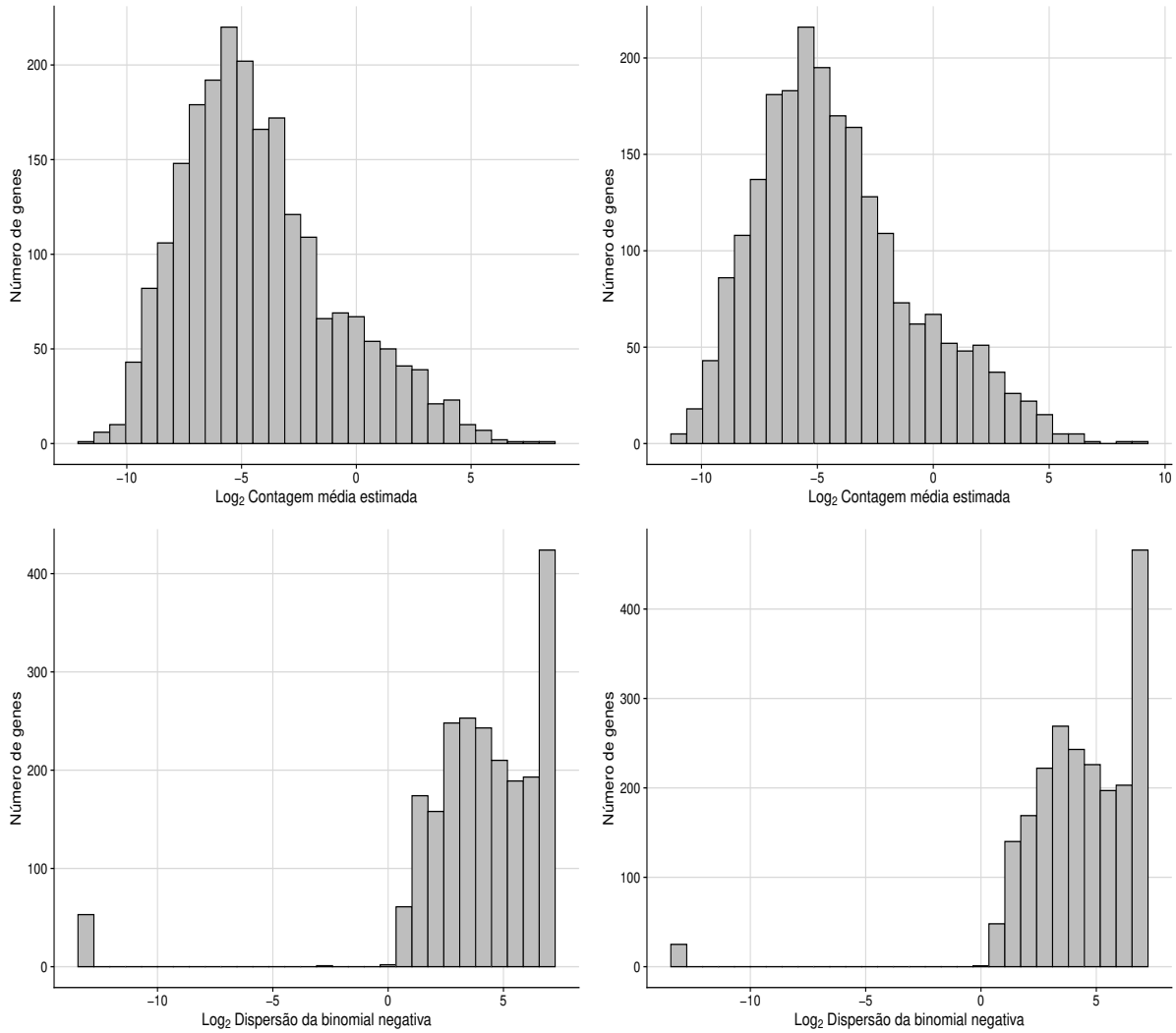


Figura 4.2 – Distribuição da contagem média estimada e parâmetro de dispersão para cada gene. Gráficos à esquerda considerando três indivíduos e à direita seis indivíduos.

Para estimar a variabilidade do efeito de indivíduo, isto é, o parâmetro σ^2, o modelo Binomial Negativa misto com efeito aleatório de indivíduo foi estimado utilizando a função `glmer` do pacote **lme4** (BATES et al., 2015). O modelo considera apenas intercepto como efeito fixo e o fator de escala foi utilizado como *offset*. A estimativa da variância do efeito aleatório de indivíduo para cada gene é utilizada como estimativa de σ^2. Entretanto, como cada gene possui uma particularidade na estimação da variância do efeito aleatório, como pode ser visto pela Figura 4.3 utilizou-se como estimativa para σ^2 a média entre todos os genes da variância do efeito aleatório de indivíduo estimada. Assim, para três indivíduos a estimativa foi $\widehat{\sigma}^2 = 0,0903$, já para seis indivíduo $\widehat{\sigma}^2 = 0,1299$.

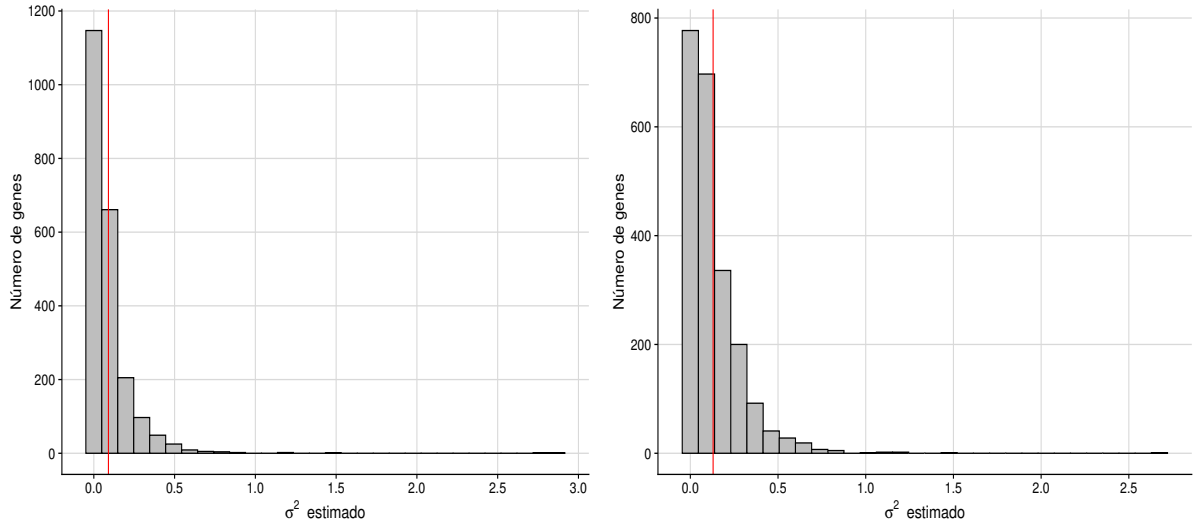


Figura 4.3 – Distribuição da variância do efeito aleatório estimada. Gráficos à esquerda considerando três indivíduos e à direita seis indivíduos.

Para melhor entendimento do processo de simulação, a Figura 4.4 apresenta o modelo empregado para simulação dos dados. Note que pela formulação do modelo o parâmetro δ_{gjk} é estimado por e^{ω_j}, enquanto que o parâmetro θ_{ijk} por k_i.

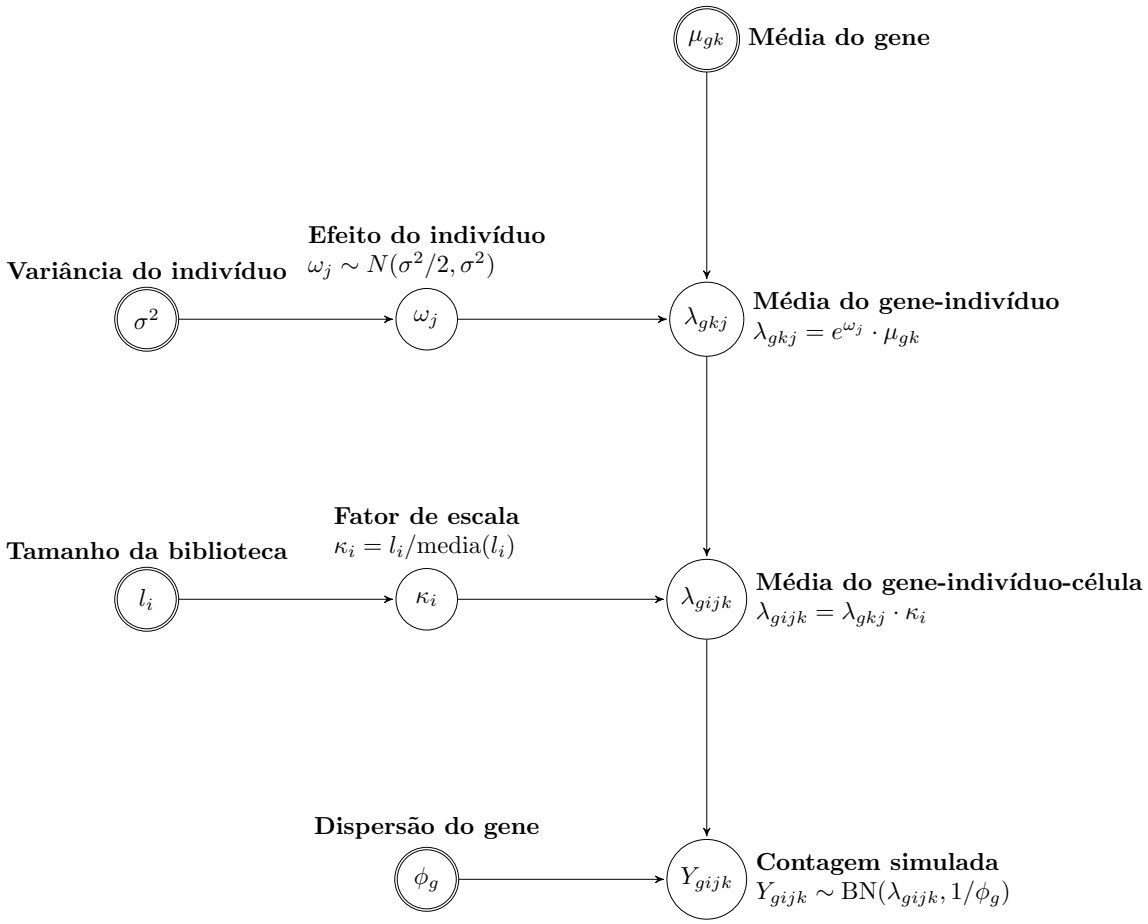


Figura 4.4 – Diagrama ilustrando o modelo de simulação utilizado. Os valores de entrada e estimados a partir dos dados são indicados pelo duplo círculo.

4.1.3 Cenários da simulação

Para reproduzir um delineamento similar ao encontrado nos dados de Liao et al. (2020) consideramos cenários com 2 e 3 grupos de condições biológicas distintas e 3 e 6 indivíduos em cada grupo. A quantidade de células para cada indivíduo em cada grupo foram amostradas de uma distribuição Uniforme discreta com suporte em $[200, 500]$. O parâmetro θ_{ijk} foi reamostrado com reposição do fator de escala estimado correspondente ao indivíduo selecionado. Além disso, 20.000 genes foram selecionados com reposição, sendo a média μ_{gk} de cada estimada a partir dos dados. Por fim, duas populações de células foram geradas, repetindo o processo de simulação de forma independente.

Os modelos considerados foram o modelo linear misto (MLM) utilizando como variável resposta o logaritmo da expressão gênica normalizada pelo fator de escala *deconvolution* (LUN; BACH; MARIONI, 2016). Neste modelo, a estatística t com graus de liberdade baseados na aproximação de Satterthwaite e os contrastes apropriados foi utilizada para avaliar diferenças entre os grupos. Além disso, o modelo edgeR proposto por

Robinson, McCarthy e Smyth (2009) foi considerado e o teste da razão de verossimilhanças com os contrastes apropriados foi conduzido para testar diferenças entre os grupos dentro de cada tipo de célula. O modelo edgeR foi empregado diretamente nas expressões gênicas das células (edgeR sc) e nas expressões das amostras pseudo *bulk* (edgeR bulk), onde as contagens foram somadas em cada nível dos grupos e tipos de células seguindo a proposta de Lun e Marioni (2017).

A Tabela 4.1 apresenta os parâmetros dos cenários utilizados para avaliar o erro do Tipo I dos modelos na análise DE. Para cada cenário a simulação foi repetida 10 vezes afim de garantir estabilidade dos resultados, a última coluna da Tabela 4.1 apresenta o total médio de células das 10 iterações para cada cenário.

Tabela 4.1 – Resumo dos cenários da simulação para avaliar erro do Tipo I.

Cenário	Efeito de indivíduo	Total de grupos	Total de indivíduos	Total médio de células
1	Sim	2	3	4.307
2	Sim	3	3	6.264
3	Sim	2	6	8.118
4	Sim	3	6	12.556
5	Não	2	3	3.750
6	Não	3	3	6.242
7	Não	2	6	8.498
8	Não	3	6	13.310

A Figura 4.5 apresenta os dados gerados para uma iteração Monte Carlo utilizando a representação UMAP. Percebe-se claramente o efeito dos indivíduos nos cenários 1 a 4, além disso, há também duas diferentes populações de células denominadas Z e Y. Por outro lado, nos cenários 6 a 8, como esperado os grupos de células estão agrupados conforme apenas o tipo de célula.

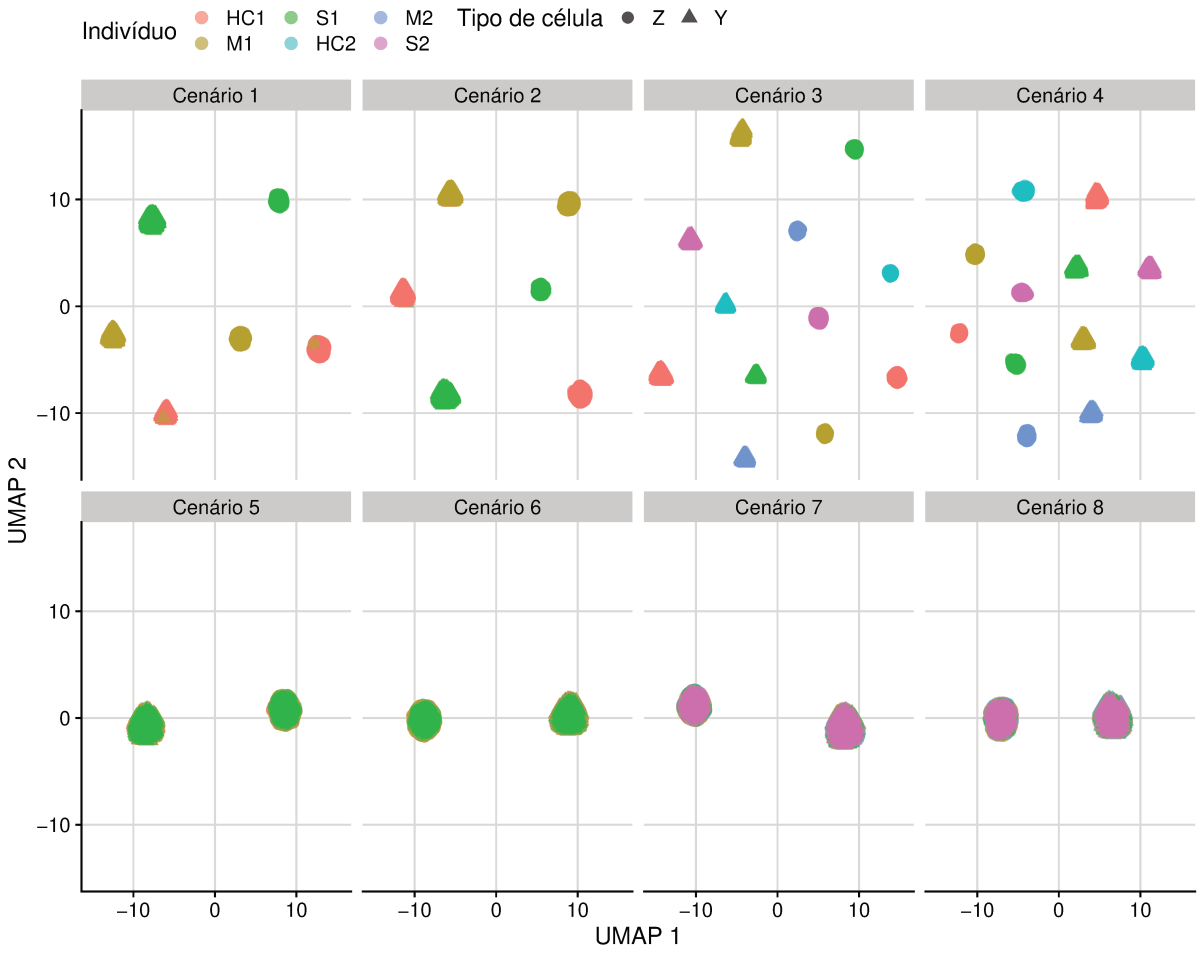


Figura 4.5 – Representação UMAP dos dados gerados para uma iteração Monte Carlo de acordo com o cenário para avaliar o erro do Tipo I dos modelos.

Para avaliar o poder e FDR dos modelos é necessário gerar genes com diferença na expressão entre os grupos. Sem perdas de generalidades, considerou-se apenas uma população de células com dois grupos biologicamente distintos e três indivíduos em cada grupo. Dos 20.000 genes simulados um conjunto aleatório foi selecionado para incorporar diferença entre os grupos, sendo esse conjunto particionado igualmente com alta e baixa regulação na diferença. A Tabela 4.2 apresenta os parâmetros utilizados em cada cenário da simulação para avaliar o poder dos modelos. Os percentuais de genes DE e o *fold change* escolhidos foram baseados em experiencias reais.

Tabela 4.2 – Resumo dos cenários da simulação para avaliar o poder.

Cenário	Efeito de indivíduo	*fold change*	% de genes DE	Total médio de células
1	Sim	0,5%	2	2.090
2	Sim	1,0%	2	2.310
3	Sim	2,0%	2	2.096
4	Sim	5,0%	2	2.285
5	Sim	0,5%	4	2.233
6	Sim	1,0%	4	2.585
7	Sim	2,0%	4	1.952
8	Sim	5,0%	4	2.239
9	Não	0,5%	2	2.413
10	Não	1,0%	2	2.033
11	Não	2,0%	2	2.053
12	Não	5,0%	2	2.111
13	Não	0,5%	4	2.163
14	Não	1,0%	4	2.367
15	Não	2,0%	4	1.811
16	Não	5,0%	4	2.354

Para ilustrar os dados gerados em cada cenário do estudo para avaliar o poder dos modelos a Figura 4.6 apresenta a representação UMAP dos dados. Observa-se o efeito de indivíduo presente nos cenários 1 a 8, sendo mais evidente a diferenças entre os grupos A e B nos cenários 7 e 8, os quais possuem *fold change* de 4 com 2% e 5% de genes DE, respectivamente. Os cenários 9 a 16 onde não há efeito de indivíduo, as células estão distribuídas de forma aleatório, como era de se esperar.

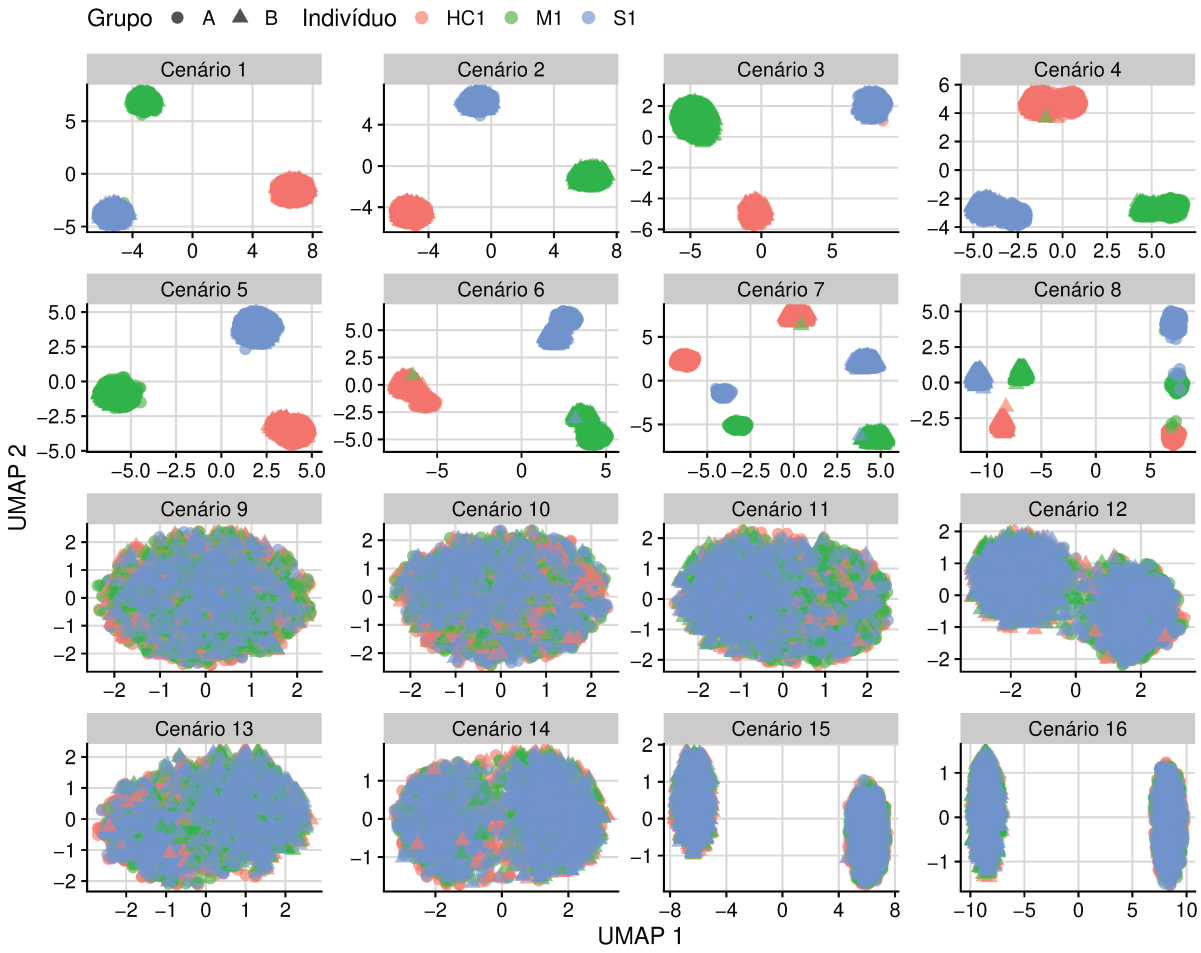


Figura 4.6 – Representação UMAP dos dados gerados para uma iteração Monte Carlo de acordo com o cenários para avaliar o poder dos modelos.

4.1.4 Resultados e discussões

Nas simulações para avaliar o erro do Tipo I (ver cenários da Tabela 4.1) a verdadeira hipótese nula para cada gene g é que μ_{gk} é igual para todos os k grupos. Assim, a rejeição da hipótese nula representa o erro do Tipo I. Para um nível de significância estabelecido α, o erro do Tipo I estimado foi definido como a proporção de todos os genes com valor-p abaixo de α. Para garantir estabilidade na estimativa, a média nas 10 iterações Monte Carlo foi utilizada, assim o erro do Tipo I estimado é definido por

$$\frac{1}{M}\frac{1}{G}\sum_{m=1}^{M}\sum_{g=1}^{G} I\left(p_{gm} \leqslant \alpha\right)$$

em que M é o número de réplicas Monte Carlo, G é o total de genes, e p_{gm} é o valor-p resultando do teste de hipótese para o gene g na réplica Monte Carlo m.

No conceitos da inferência clássica dos testes de hipótese um teste estatístico é considerado exato se a taxa de erro do Tipo I é igual a α. Por outro lado, um teste é conservativo quando a taxa de erro do Tipo I é não excede o valor α. Além disso,

classifica-se um teste liberal quando a taxa de erro do Tipo I é superior a α (ROHATGI, 1984).

Os resultados da simulação estão apresentados na Figura 4.7. Percebe-se a taxa de erro do Tipo I estimada quando há efeito de indivíduo excede o nível estabelecido de 5% para o modelo edgeR ajustado diretamente nas amostras observadas (edgeR sc) e o modelo linear misto (MLM). Entretanto, a discrepância entre o nível especificado e estimado é maior para o modelo edgeR sc, em torno de 0,20 e 0,25 para os cenários com 3 e 6 indivíduos, respectivamente. O MLM excede o nível estabelecido em torno de 0,08 para todos os cenários. Essa característica liberal dos modelos edgeR sc e MLM é atribuída a presença do efeito de indivíduo. Quando a simulação é repetida sem o efeito do indivíduo observa-se que as taxas do erro do Tipo I são substancialmente menores e mais próximas do nível estabelecido. Dessa forma, os resultados sugerem que as análise DE terão um desempenho errôneo se o efeito de indivíduo for ignorado, especialmente para o modelo edgeR. A perda do controle de erro do tipo I resultará em um número inaceitável de falsos positivos no conjunto final de genes DE. Pelo fato do MLM capturar diretamente o efeito do indivíduo as taxas de erro do Tipo são abaixo do modelo edgeR, porém acima acima do nível estabelecido.

Analisando os resultados da abordagem edgeR *bulk*, isto é, considerando amostras pseudo *bulk*, verifica-se que a taxa do erro do Tipo I está muito abaixo do nível estabelecido, como valores em torno de 0,004 e 0,003 quando há efeito de 3 e 6 indivíduos, respectivamente. A característica conservadora da abordagem *bulk* pode ser interessante na análise de dados reais, uma vez que resultará em um número abaixo do nominal de falsos positivos.

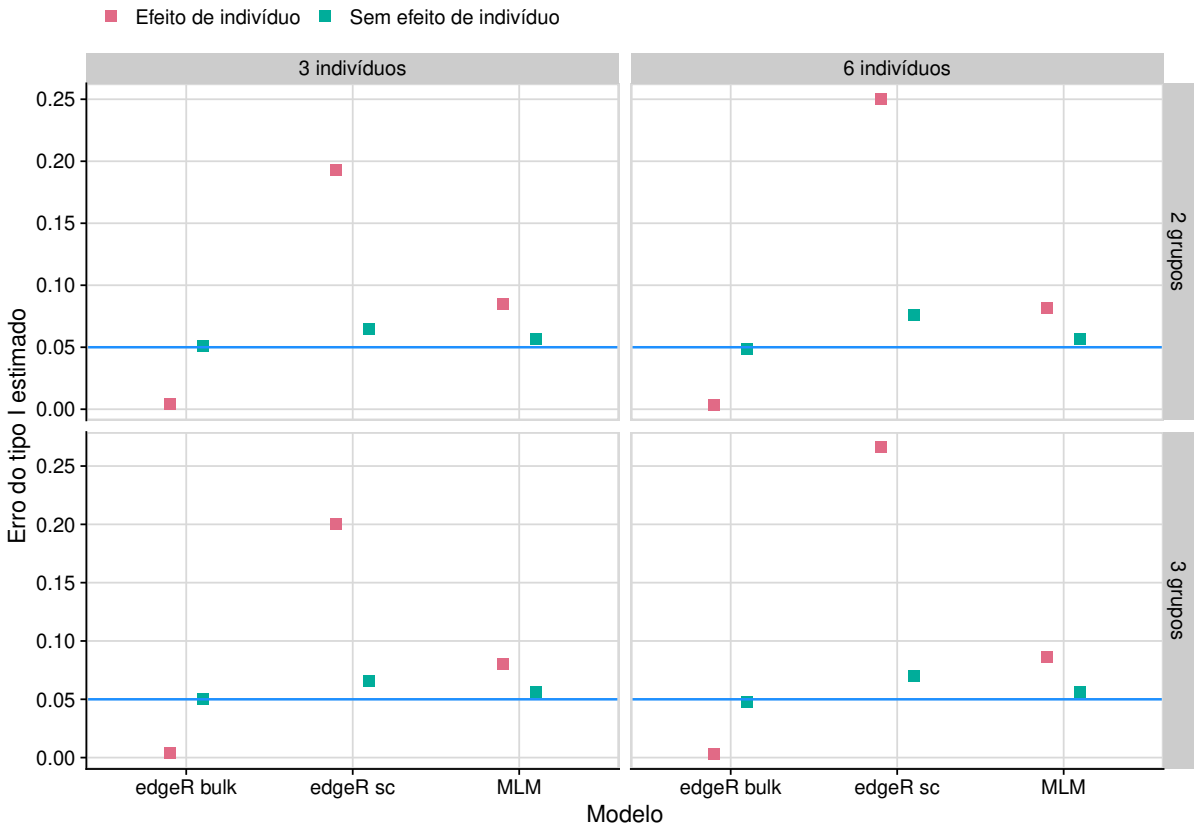


Figura 4.7 – Erro do Tipo I estimado considerando $\alpha = 5\%$ dos modelos de acordo com o cenário. Cada erro do Tipo I estimado representa a média de 10 iterações Monte Carlo.

Seguindo os cenários discutidos da Tabela 4.2, genes DE foram incorporados na análise e a taxa de falsas descobertas (*false discorvery rate* – FDR) e o poder estatístico dos métodos foram avaliados. A Figura 4.8 apresenta o FDR estimado considerando nível estabelecido de $\alpha = 5\%$. Observa-se que o MLM controla de forma exata ou abaixo do nível especificado o FDR quando há efeito de individuo. Em contrapartida, o modelo edgeR sc não consegue controlar o FDR na presença de efeito de indivíduo, com exceção do cenário com 5% de genes DE. A estratégia utilizando amostras *bulk* apresenta ótimo controle do FDR, com valores abaixo do nominal, para os cenários onde há efeito de indivíduo. De forma geral, à medida que aumenta a proporção de genes DE os métodos se tornam mais acurados no controle do FDR.

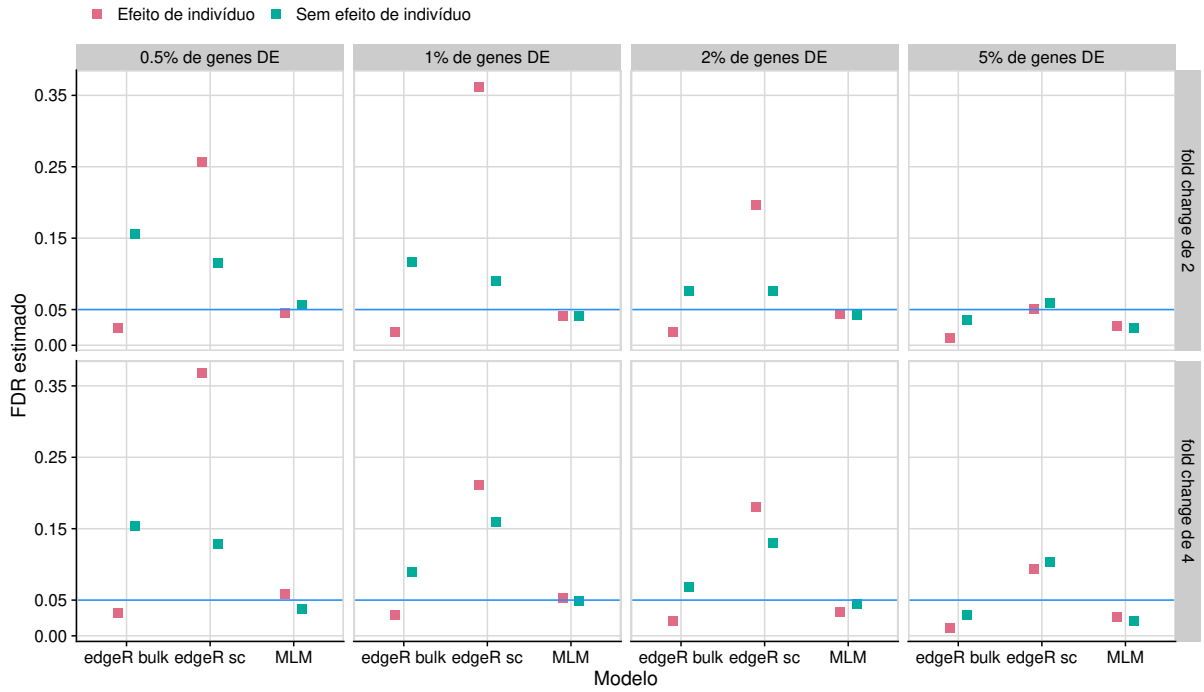


Figura 4.8 – FDR estimado considerando $\alpha = 5\%$ dos modelos de acordo com o cenário. Cada FDR representa a média de 10 iterações Monte Carlo.

A curva ROC e sua área (AUC) são métricas utilizadas por vários autores, por exemplo McCarthy e Smyth (2009), Lun e Marioni (2017) e Soneson e Robinson (2018) na avaliação do poder dos métodos. Em particular, a AUC indica se o método é capaz de detectar os verdadeiros genes expressamente diferentes em relação aqueles genes que não são DE. Os resultados apresentados na Figura 4.9 indicam que os modelos edgeR tem AUC próxima de 0,99, enquanto que o MLM apresenta AUC entre 0.94 e 0.95. De forma geral, a medida que aumenta o percentual de genes DE maior o poder dos testes. Além disso, quando o *fold change* é de 4 para os genes DE todos os métodos apresentam AUC próximas de um.

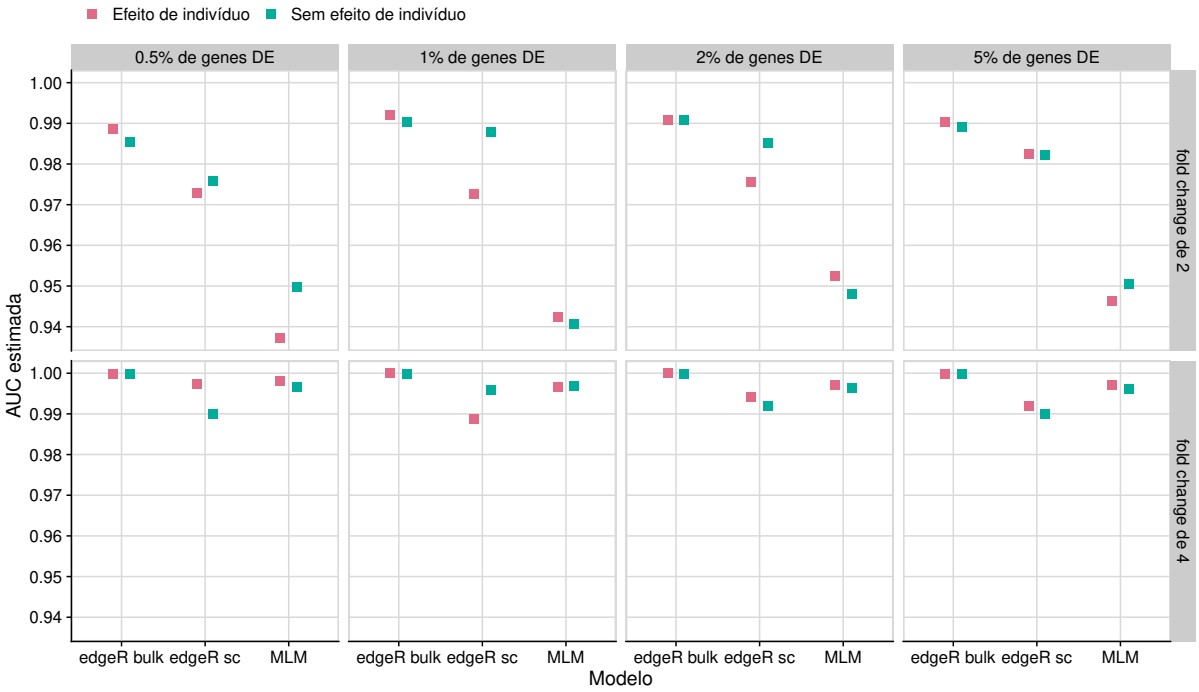


Figura 4.9 – Área sob a curva ROC (AUC) dos modelos de acordo com o cenário. Cada AUC representa a média de 10 iterações Monte Carlo.

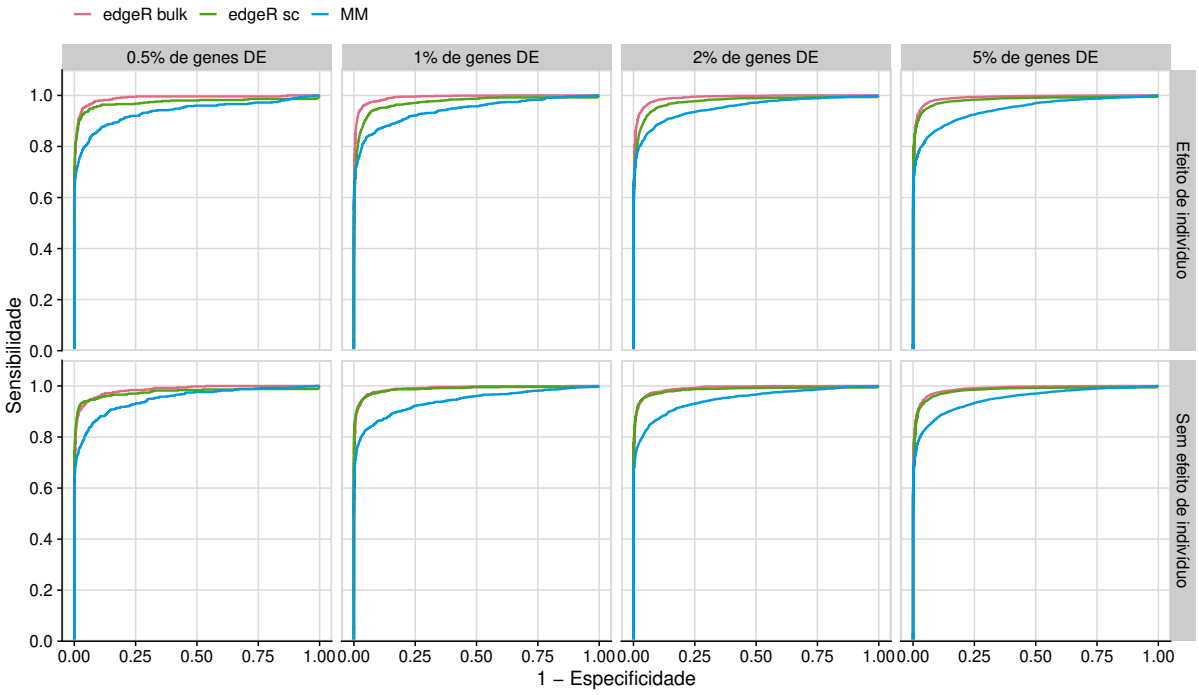


Figura 4.10 – Curva ROC dos modelos de acordo com o cenário. Cada curva representa a média de 10 iterações Monte Carlo.

Baseado nos estudos de simulação realizados concluí-se que o modelo linear misto considerando como varável resposta o logaritmo das contagens normalizadas pelo método

deconvolution apresentou desempenho satisfatório, tendo em vista, que a distribuição da variável resposta condicionada nos grupos não é Normal, justificando a característica liberal no controle do erro do Tipo I relatada na Figura 4.7. Este estudo vem de encontro com os resultados apresentados por Lun e Marioni (2017), onde os autores também mostram que as técnicas usuais para análise DE não controlam o erro do Tipo I na presença de fatores aleatórios e a abordagem utilizando amostras pseudo *bulk* elimina o efeito aleatório presente nas células controlando tanto o erro do Tipo I quanto o FDR sem a perda do poder dos testes. Portanto, é fundamental incorporar o efeito do indivíduo na análise DE.

4.2 Análise das células BALF em indivíduos com COVID-19

O coronavírus da síndrome respiratória aguda grave 2 (SARS-CoV-2) é um vírus de RNA que surgiu em Dezembro de 2019 em Wuhan, China e é a causa da doença coronavírus 2019 (COVID-19), a qual existe até o presente momento uma pandemia em curso. Liao et al. (2020) realizaram uma análise utilizando dados de um experimento de scRNA-seq em células do líquido de lavagem broncoalveolar (BALF) de doze indivíduos nas seguintes condições de saúde

- três pacientes com COVID-19 em estado moderado (M1-M3);
- seis pacientes com COVID-19 em estado grave (S1-S6);
- três pacientes controle (HC1-HC3).

Após coleta do material biológico e sequenciamento para cada indivíduo as amostras foram lidas e no utilizando o programa *Cell Ranger Software Suite*, versão 3.1.0 e alinhadas ao genoma de referência. Segundo os autores, para detectar leituras do vírus SARS-CoV-2, uma referência personalizada foi construída integrando o genoma humano GRCh38 e SARS-CoV-2, disponível no *National Center for Biotechnology Information* (NCBI) pelo código MN908947.3. As matrizes com as contagens UMI para cada indivíduo estão disponíveis no repositório público *Gene Expression Omnibus* (GEO) pelo código de identificação GSM4339769.

Liao et al. (2020) conduziram a análise seguindo o fluxo usual presente no pacote Seurat versão 3 (STUART et al., 2019). Especificamente, a analise de agrupamento foi realizada considerando métodos baseados em grafos e a analise de diferenciação na expressão foi conduzida empregando o modelo MAST proposto por Finak et al. (2015).

Nesta seção, o objetivo é realizar o fluxo usual seguindo a metodologia presente discutida por Amezquita et al. (2020) e disponível principalmente nos pacotes do Bioconductor. O processamento das matrizes de contagem é conduzido seguindo as abordagens discutidas no Capítulo 2, a análise de agrupamento é realizada utilizando o método *k-means*, sendo a escolha do número de grupos e a estabilidade dos grupos avaliada seguindo a metodologia apresentada no Capítulo 3, além disso, a anotação biológica e detecção de gene marcadores também é realizada a fim de fornecer interpretabilidade aos grupos encontrados. Finalmente, a análise de expressão gênica é conduzida utilizando modelos que incorporam o efeito de indivíduo presente nos dados.

4.2.1 Processamento dos dados

Nesta primeira seção o usual fluxo de análise de dados de scRNA-seq foi empregado com intuito de caracterizar as células BALF dos doze indivíduos. O comportamento

das células de cada indivíduo é apresentado na Tabela 4.3.

Tabela 4.3 – Características de cada amostra coletada.

ID	Total de células	Tamanho biblioteca	Total de genes	Genes iguais a zero	Genes na mitocôndria	Tamanho da mitocôndria
HC1	11.115	47.013.233	33.538	13.988 (41,7%)	13	1.209.895 (2,6%)
HC2	10.366	26.522.426	33.538	15.010 (44,8%)	13	975.360 (3,7%)
HC3	8.972	82.693.340	33.538	12.301 (36,7%)	13	16.129.222 (19,5%)
M1	6.249	67.351.642	33.539	10.080 (30,1%)	13	4.616.789 (6,9%)
M2	10.269	70.961.621	33.539	10.373 (30,9%)	13	10.756.486 (15,2%)
M3	3.716	13.315.819	33.539	14.007 (41,8%)	13	8.621.064 (64,7%)
S1	18.044	47.995.033	33.539	10.417 (31,1%)	13	2.102.881 (4,4%)
S2	20.857	124.908.876	33.539	9.166 (27,3%)	13	5.350.913 (4,3%)
S3	4.111	10.142.512	33.539	11.748 (35,1%)	13	1.868.949 (18,4%)
S4	3.920	28.875.262	33.539	10.710 (31,9%)	13	3.285.438 (11,4%)
S5	2.879	15.127.794	33.539	11.902 (35,5%)	13	995.567 (6,6%)
S6	7.732	29.214.231	33.539	10.502 (31,3%)	13	3.055.347 (10,5%)

Nas Figuras 4.11, 4.12, 4.13 e 4.14 cada ponto representa uma célula e está colorido conforme o critério de baixa qualidade definido por Liao et al. (2020), em particular, se alguma dessas condições forem satisfeitas:

- Tamanho da biblioteca: maior que 1.000.
- Total de genes detectados: entre 200 a 6.000.
- Proporção de genes mitocondriais: menor que 0,1.

De forma geral, os resultados indicam que a maioria das células apresentam alta expressão gênica de genes mitocondriais, tal fenômeno é um indicativo de células com baixa qualidade, possivelmente por causa da perda de RNA citoplasmático das determinadas células (ISLAM et al., 2014). Como diagnóstico para garantir que células de alta qualidade que são metabolicamente ativas não sejam removidas a Figura 4.14 aponta que não há células no canto superior direito, que seria um indício de células metabolicamente ativas e não danificadas.

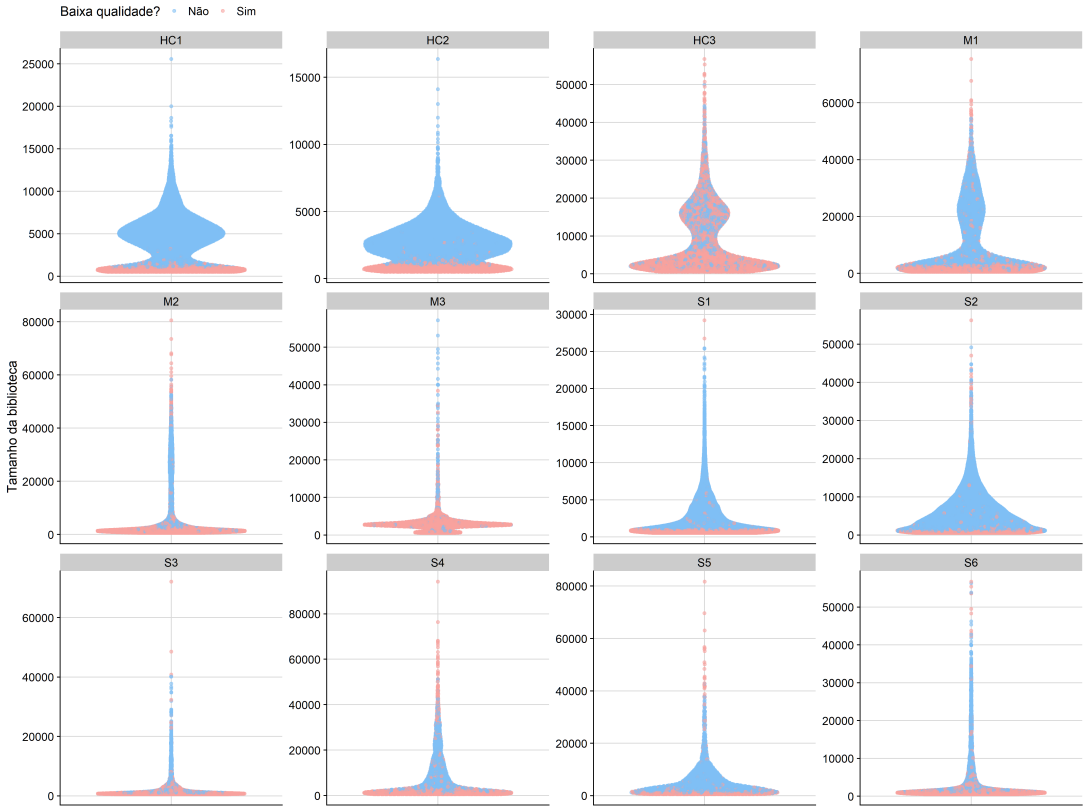


Figura 4.11 – Distribuição do tamanho da biblioteca conforme o indivíduo.

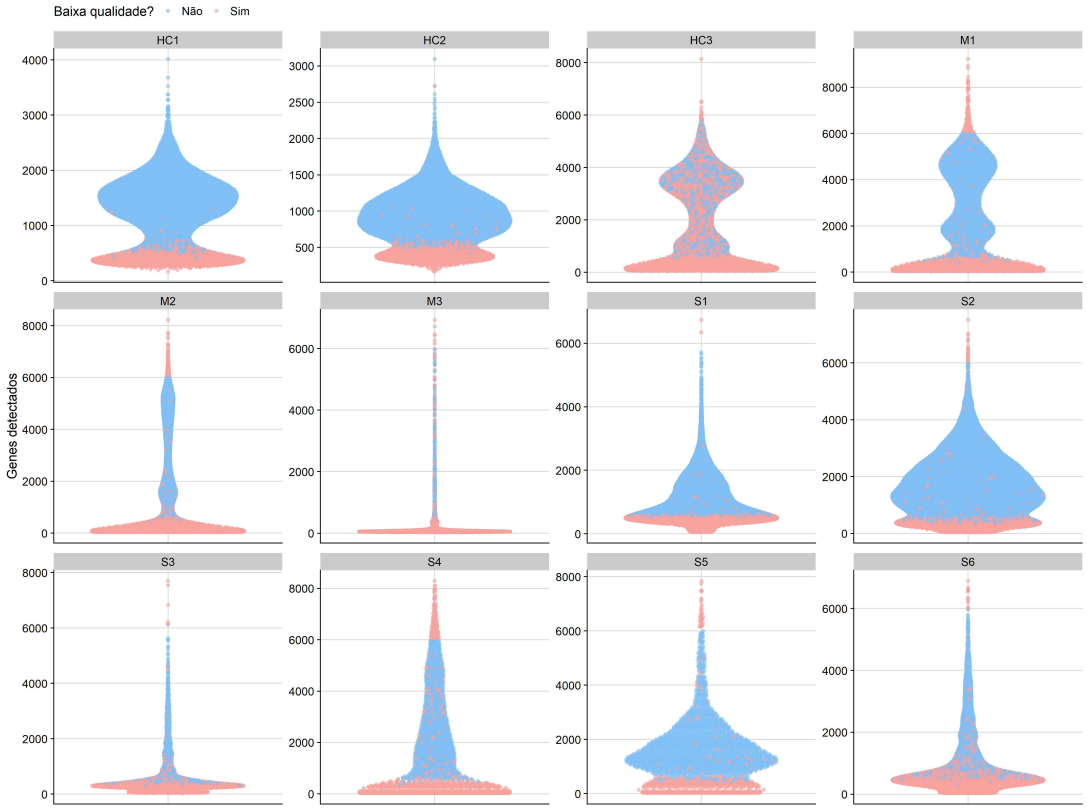


Figura 4.12 – Distribuição do total de genes detectados por célula conforme o indivíduo.

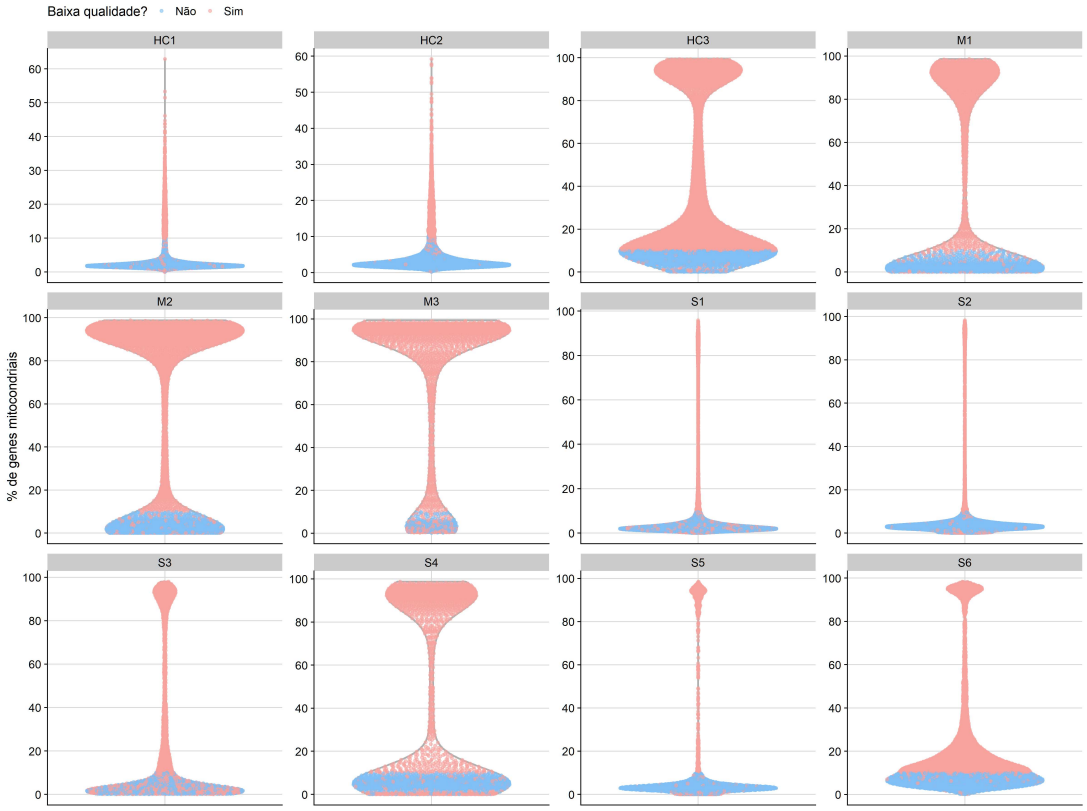


Figura 4.13 – Distribuição da proporção de genes mitocôndrias detectados em cada célula conforme o indivíduo.

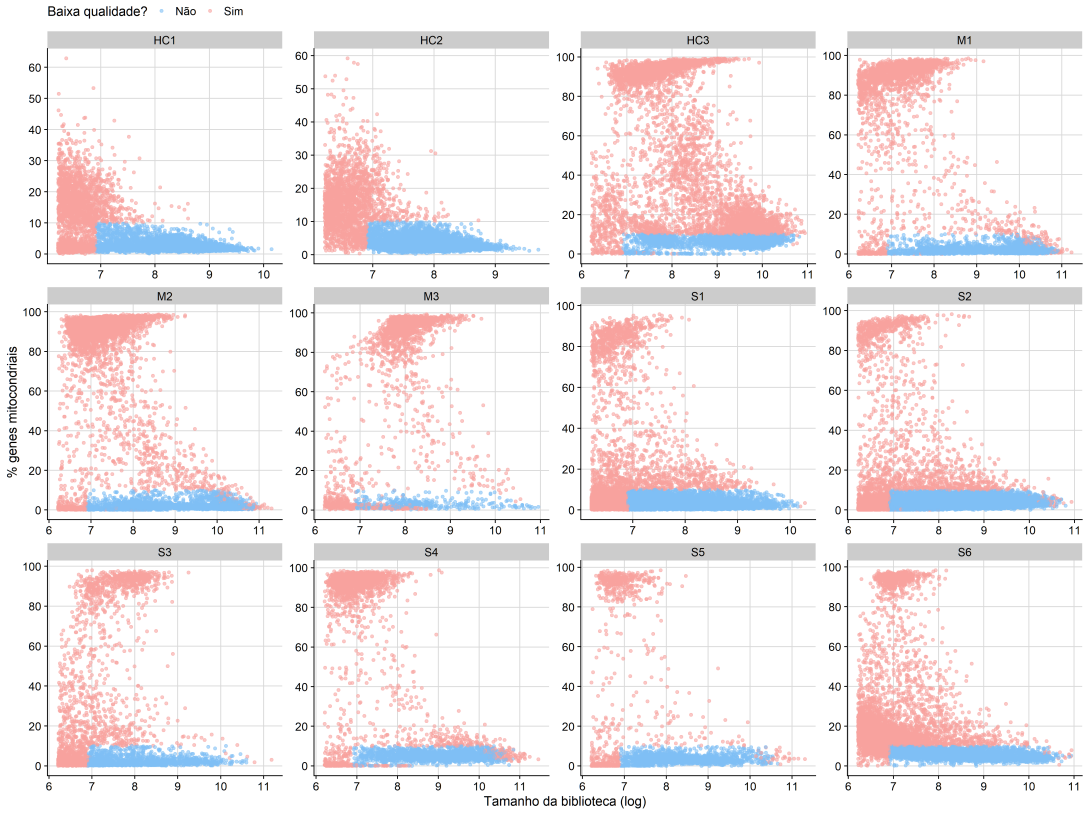


Figura 4.14 – Percentual de genes mitocondriais versus tamanho da biblioteca conforme o indivíduo.

Seguindo o critério adotado a Tabela 4.4 apresenta o resumo da quantidade de células descartadas conforme o critério para cada indivíduo do estudo. Analisando os resultados podemos observar que a proporção de genes mitocondriais foi o critério que maior detectou células de baixa qualidade. Interessante mencionar também que 90,18% das células do indivíduo M3 foram consideradas de baixa qualidade. No nível das *features* os genes com expressão igual a zero para todos os indivíduos foram removidos da análise. Assim, dos 33.538 genes presentes em todos os indivíduos restaram-se 14.240 (42%).

Tabela 4.4 – Resumo das células descartadas conforme o critério para cada indivíduo.

ID	Tamanho biblioteca	Genes detectados	Genes mitocondriais	Total de descarte	% de descarte
HC1	2.340	3	1.510	2.648	23,82
HC2	1.842	5	1.661	2.174	20,97
HC3	555	2.397	6.270	6.405	71,39
M1	840	2.131	2.418	2.707	43,32
M2	1.413	5.826	6.413	6.857	66,77
M3	333	2.807	2.960	3.351	90,18
S1	5.331	650	1.729	6.164	34,16
S2	2.531	594	1.529	3.514	16,85
S3	1.856	716	1.295	2.812	68,40
S4	1.010	1.647	1.787	2.202	56,17
S5	513	426	534	808	28,07
S6	2.387	783	4.409	4.829	62,45

Após o controle de qualidade, as contagens observadas de cada gene foram normalizadas e transformadas na escala logarítmica de base 2. A normalização foi realizada para cada indivíduo de forma independente considerando o pelo método *deconvolution* (LUN; BACH; MARIONI, 2016). A Figura 4.15 compara o fator de escala utilizado na normalização obtido pelo método *deconvolution* e a abordagem usual que consiste em utilizar o tamanho da biblioteca (soma das contagens por células).

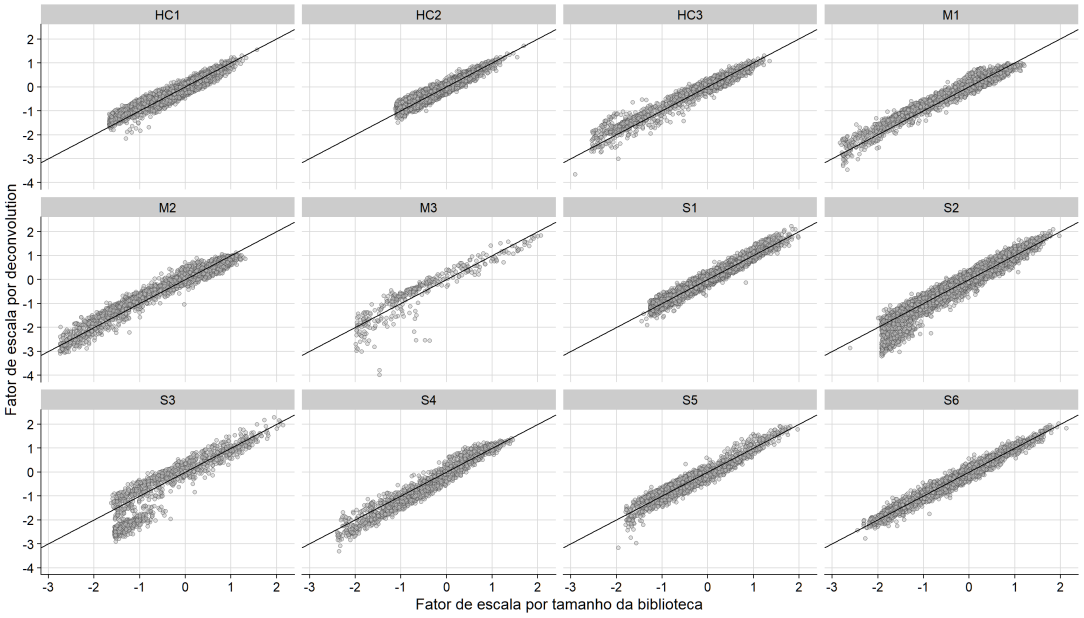


Figura 4.15 – Comparação entre os fatores de escala tamanho da biblioteca e obtido pelo método *deconvolution* conforme o indivíduo. A linha preta corresponde à igualdade entre os fatores de escala.

A Figura 4.16 exibe o comportamento das expressões gênicas para determinados genes após normalização dos dados pelo tamanho da biblioteca e aplicação da transformação logarítmica. Observa-se que, para determinados indivíduos, a distribuição da expressão gênica é simétrica, enquanto que para outras há uma forte assimetria à direita e também comportamentos bimodais, como por exemplo para o gene FTL nos indivíduos M1, M2, S4 e S5 a expressão gênica normalizada possui claramente duas modas.

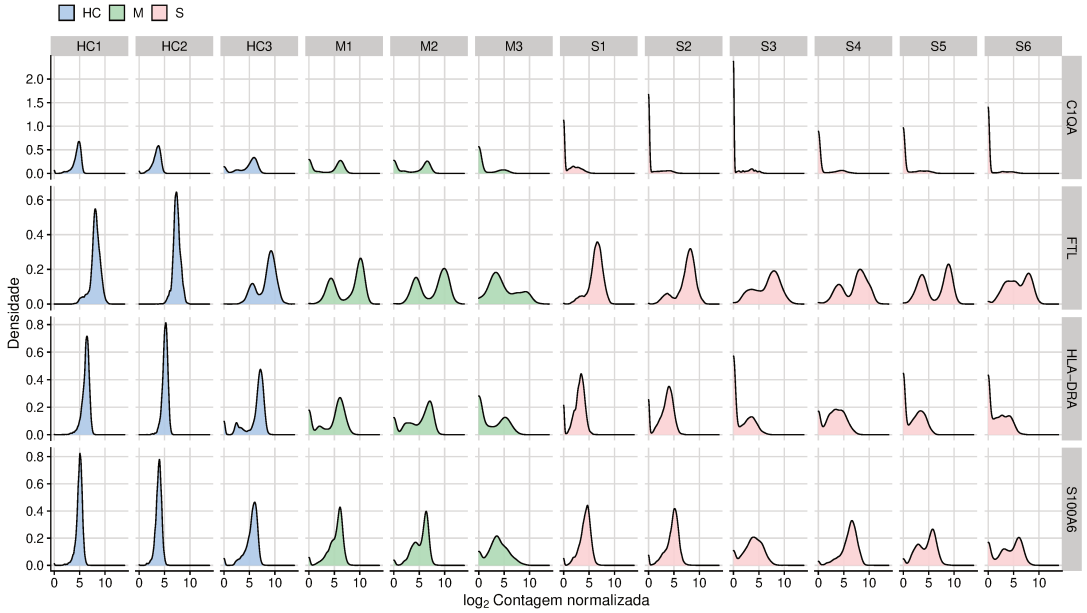


Figura 4.16 – Densidades do logaritmo da expressão gênica normalizada de alguns genes conforme o indivíduo.

Tendo em vista que existe um número expressivo de genes (14.240) e usualmente é conhecido que a expressão desses genes não são de interesse biológico, por exemplo, o organismo humano tem aproximadamente 20.000 genes sendo cerca de 0.01% diferentemente expressos em diferentes condições experimentais (ALBERTS et al., 2015), então procedeu-se com a seleção de *features*, que consiste em realizar uma redução inicial do número de genes buscando selecionar aqueles que apresentam expressão gênica devidamente a fatores biológicos.

Essa análise foi conduzida modelando para cada indivíduo a variância de cada gene em função da média de sua expressão, conforme descrito em detalhes no Capítulo 3. A decomposição da variância de cada gene foi realizada inicialmente de forma separada pois a relação variância-média é diferente de acordo com a amostra, conforme observa-se na Figura 4.17, em que a linha azul representa o ajuste do modelo, isto é, o componente técnico de variação de cada gene, assumindo que a maioria dos genes exibem um baixo nível de variação, em torno da média, que não é biologicamente interessante.

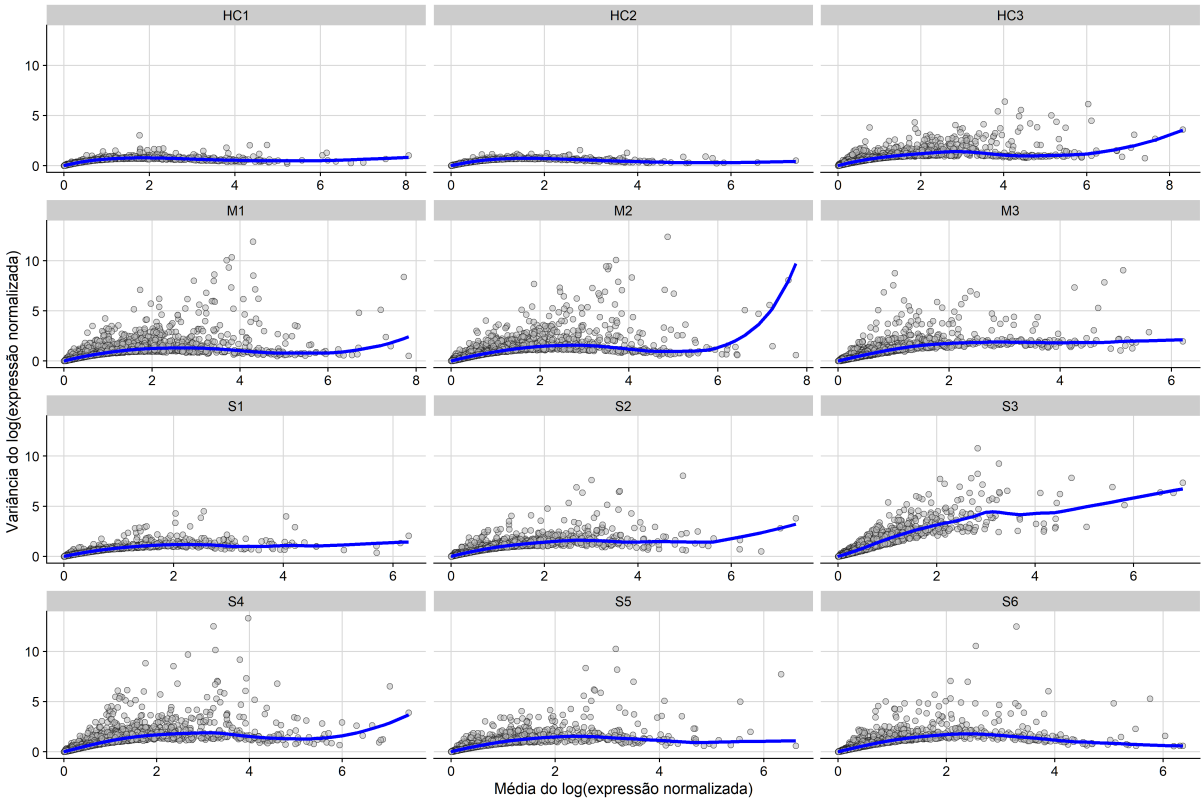


Figura 4.17 – Relação média e variância dos genes conforme o indivíduo. Cada ponto representa um gene e a curva em azul representa o modelo ajustado.

Para realizar a seleção das *features* os resultados de cada ajuste dos modelos foram sumarizados utilizando a abordagem implementada na função `combineVar` do pacote **scran** (LUN; MCCARTHY; MARIONI, 2016). A média, variância, componente técnico (valor ajustado do modelo) e componente biológico (resíduo) foram combinados considerando a média entre os indivíduos. Além disso, para cada gene testa-se a hipótese nula de que a variância é igual a média. O valor-p obtido é corrigido utilizando a abordagem de BH (FDR), pois o mesmo teste é conduzido para vários genes. O FDR de cada gene é combinado entre os indivíduos utilizando o método de Fisher (FISHER, 1925). Finalmente, a seleção das *features* é concluída considerando os genes cujo FDR combinado é menor que 5%, resultando na seleção de 2.209 genes, 15% dos 14.240 genes iniciais. A Tabela 4.5 exibe as estatísticas e o valor-p corrigido (FDR) dos 10 primeiros genes ordenados pela componente biológica (maior resíduo).

Tabela 4.5 – Estatísticas combinadas da seleção de *features* para os 10 primeiros genes ordenados pela componente biológica.

Gene	Média	Variância	Ajustado (técnico)	Resíduo (biológico)	FDR
CCL2	1.8574	5.0498	0.8944	4.1554	0.00E+00
APOC1	2.8133	4.6097	0.8823	3.7274	0.00E+00
S100A8	2.0502	4.8165	1.4522	3.3644	4.61E-246
CXCL10	1.5648	4.2759	0.9825	3.2934	0.00E+00
IFI27	2.6348	4.5527	1.3825	3.1702	0.00E+00
TYROBP	3.5249	4.5419	1.4056	3.1363	0.00E+00
CTSL	2.4678	4.5213	1.4148	3.1065	9.36E-265
C1QB	2.1429	3.9537	0.9349	3.0188	0.00E+00
APOE	2.0529	3.8825	0.9150	2.9675	0.00E+00
C1QC	1.6861	3.8985	1.0273	2.8712	0.00E+00

Um fenômeno que ocorre quando amostras do mesmo tecido de células são sequenciadas sob diferentes circunstâncias, tempo ou laboratório, é o efeito de lote (*batch effect*). O processamento dessas sequencias está, em geral, sujeito a diferenças não controláveis, como por exemplo, mudanças no operador, diferenças na qualidade do reagente, entre outras. Assim diferenças sistemáticas na expressão gênica é observada nas células de acordo com o lote. Esse fenômeno pode ser o principal motivo da heterogeneidade dos genes, escondendo as diferenças biológicas relevantes e complicando a interpretação dos resultados. Outra principal motivação para realizar a correção do efeito de lote é permitir caracterizar a heterogeneidade da população das células de uma maneira consistente entre os indivíduos.

No que tange os dados da análise, o RNA das células de cada paciente foram sequenciadas sob diferentes condições experimentais, as quais estão detalhadas em (LIAO et al., 2020). Empiricamente, observa-se nos gráficos das Figuras 4.18, 4.19 e 4.20 que as células estão fortemente separadas conforme o indivíduo. Na literatura existem vários métodos para correção do efeito de lote, uma revisão e comparação extensiva dos métodos é apresentada por Tran et al. (2020). Para correção do efeito de indivíduo na expressão gênica adotamos a abordagem implementada na função `rescaleBatches` do pacote **batchelor** que consiste em reescalar o logaritmo da expressão normalizada de cada gene utilizando a menor média entres os indivíduos. Observa-se nas Figuras 4.18, 4.19 e 4.20 que após a correção (painel direito) as células, especialmente dos indivíduos controle, HC1, HC2 e HC3, estão menos agrupadas.

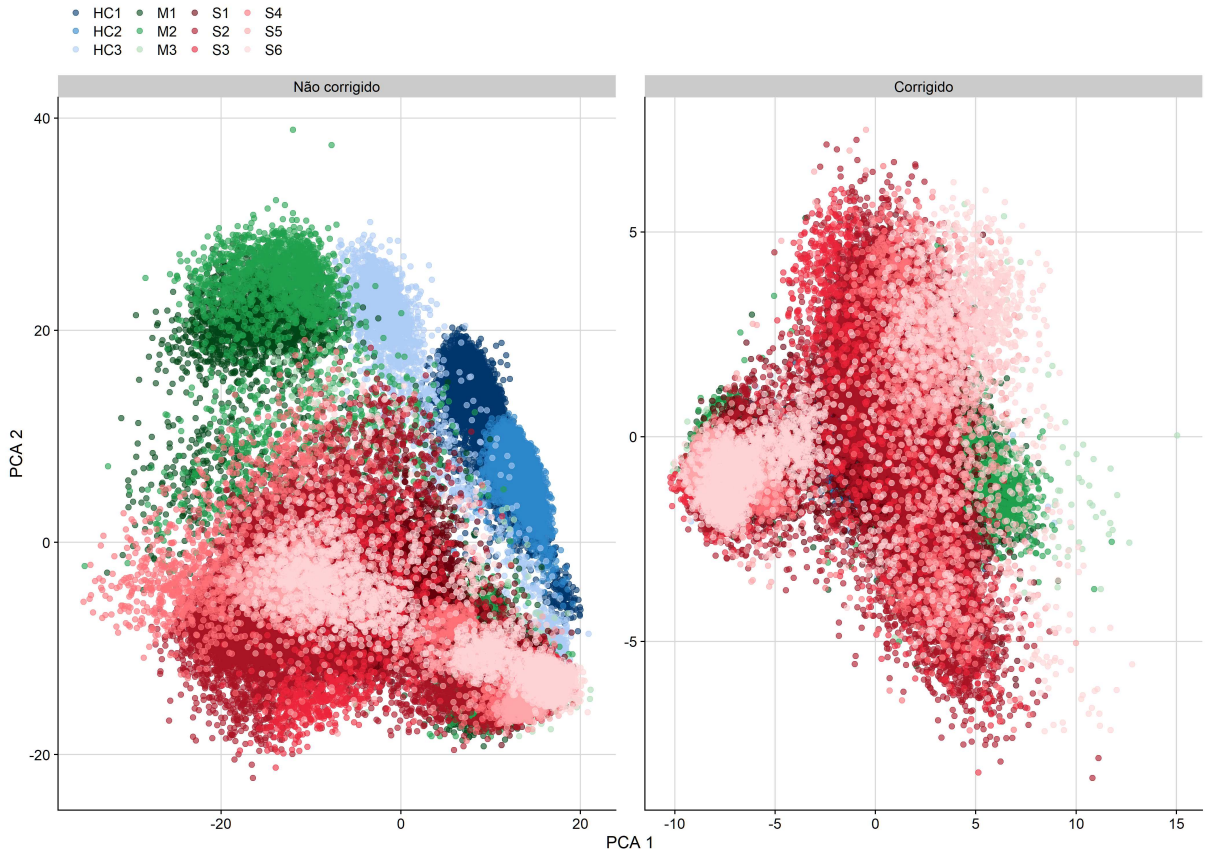


Figura 4.18 – Gráfico das duas primeiras componentes principais sem e com correção do efeito *batch* de indivíduo. Cada ponto representa uma célula e está colorido conforme o indivíduo que a mesma pertence.

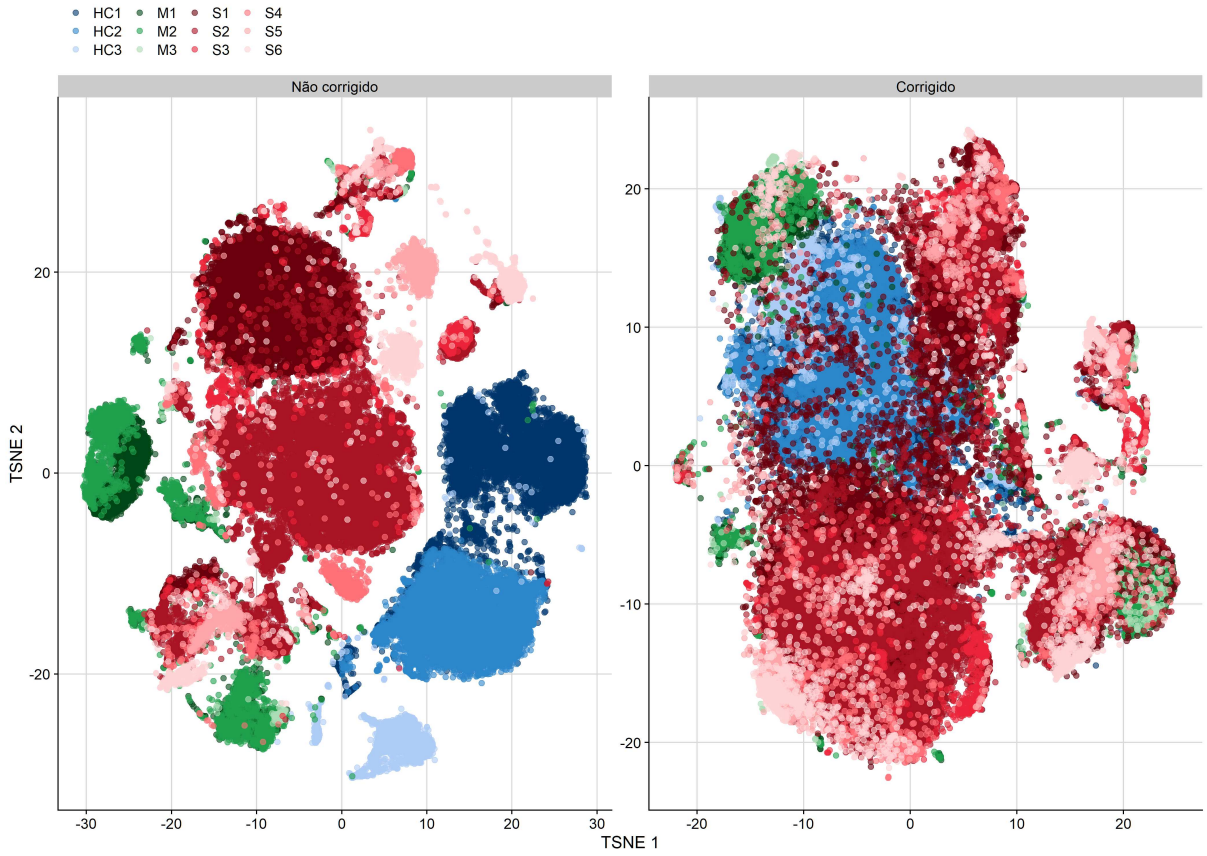


Figura 4.19 – Gráfico das duas primeiras componentes do t-SNE sem e com correção do efeito *batch* de indivíduo. Cada ponto representa uma célula e está colorido conforme o indivíduo que a mesma pertence.

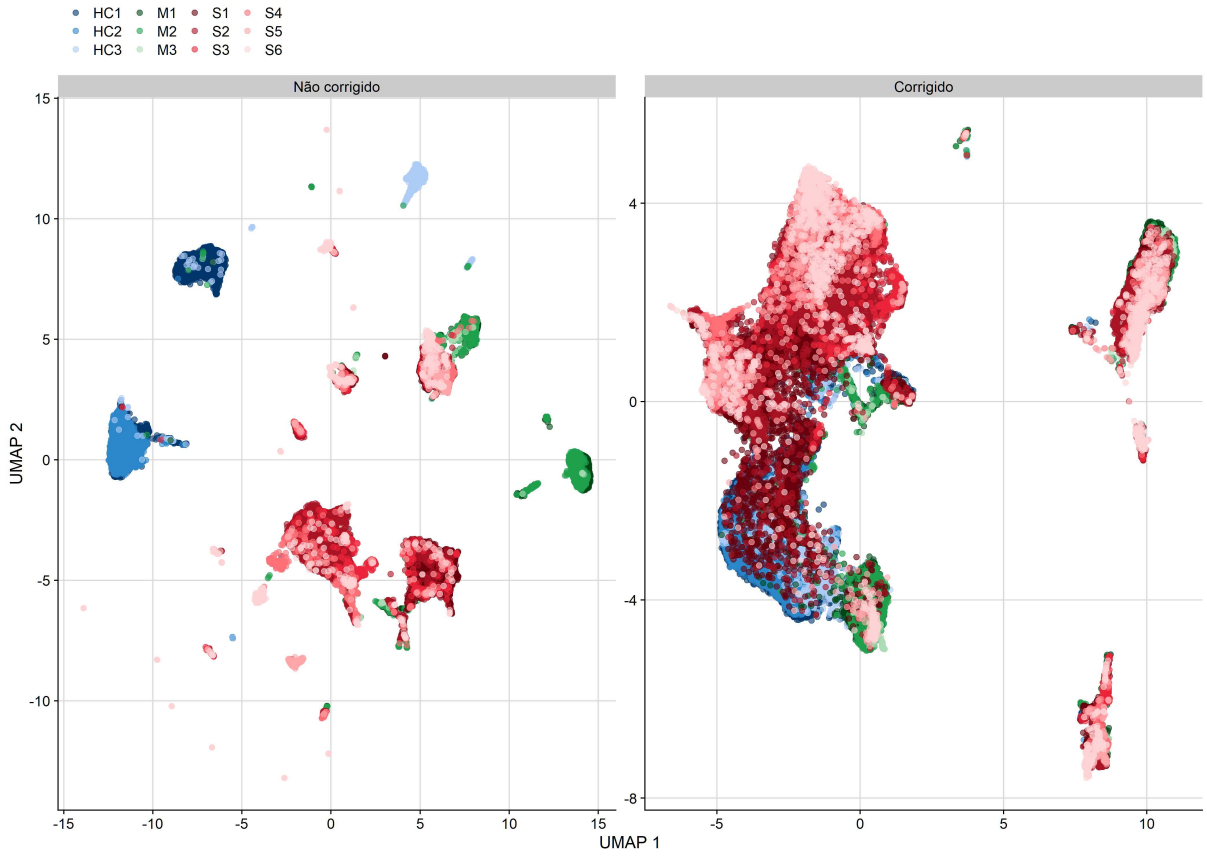


Figura 4.20 – Gráfico das duas primeiras componentes do UMAP sem e com correção do efeito *batch* de indivíduo. Cada ponto representa uma célula e está colorido conforme o indivíduo que a mesma pertence.

Para exemplificar o efeito de lote devido ao indivíduo na expressão gênica a Figura 4.21 apresenta gráficos de densidades das contagens de UMI observadas, as contagens log-normalizadas pelo método *deconvolution* e as contagens log-normalizadas e corrigidas pelo efeito de lote. Analisando as contagens de UMI verifica-se que poucas células tiveram altas contagens, enquanto que a maioria das células apresentaram contagens baixas ou iguais a zero. Por outro lado, observando o comportamento das expressões log-normalizadas é perceptível uma forte diferença das distribuições entre os indivíduos, fato esse que pode ser justificado como efeito de lote devido ao sequenciamento do indivíduo. Essa diferença é mais expressa nos indivíduos controle (HC). Em contrapartida, analisando as distribuições das expressões log-normalizadas e corrigidas verifica-se que embora haja diferenças entre indivíduos a intensidade é menor, portanto espera-se que a partir das expressões corrigidas as possíveis diferenças encontradas entre os tipos de indivíduos seja meramente biológicas e não técnica.

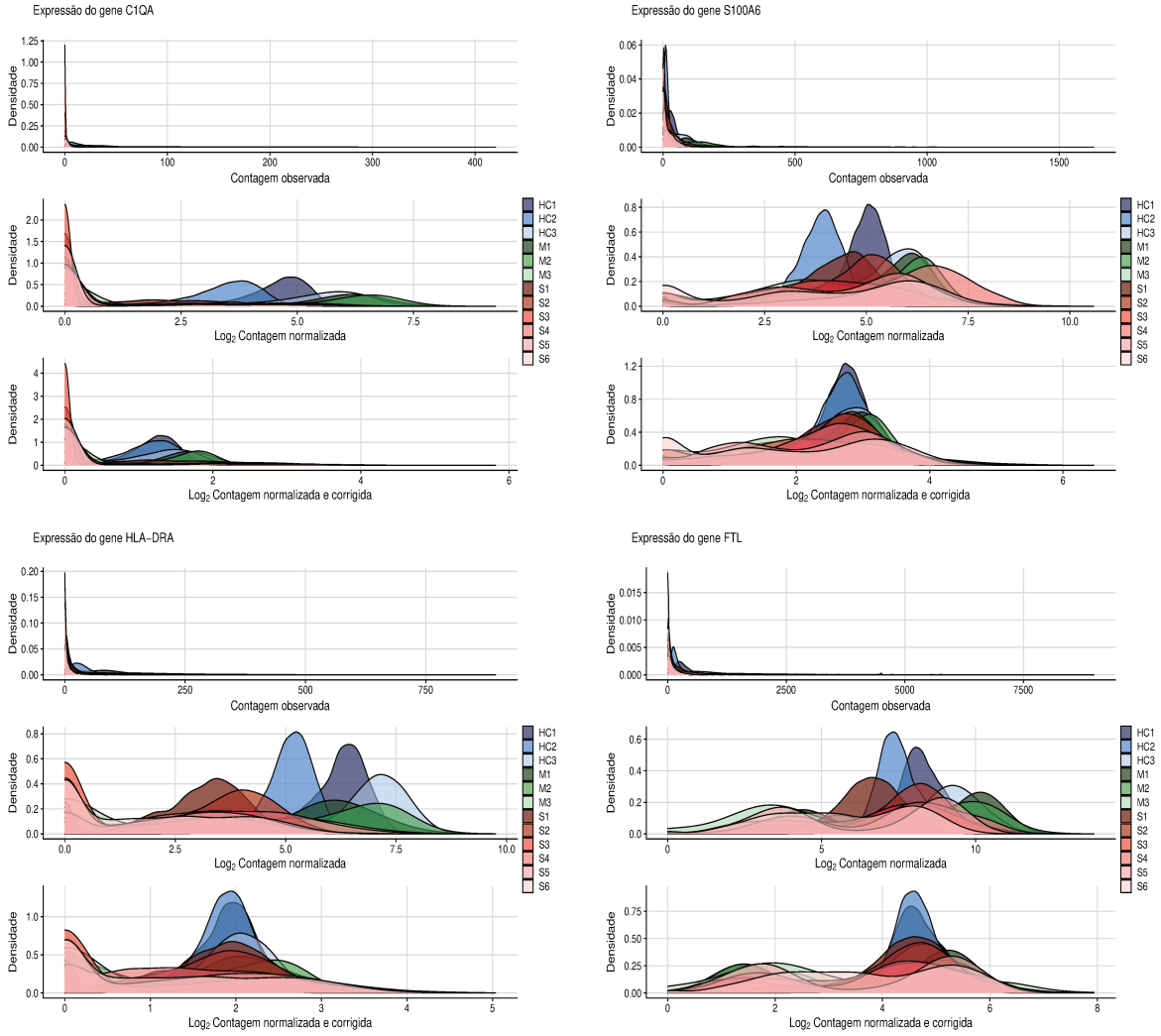


Figura 4.21 – Densidades das expressões gênicas medida em contagens UMI, log-normalizadas e log-normalizadas e corrigidas pelo efeito de lote de indivíduo para os genes C1QA, S100A6, HLA-DRA e FTL.

4.2.2 Análise de agrupamento das células BALF

Para capturar a usual heterogeneidade da expressão gênica presente em dados de scRNA-seq a análise de agrupamento é uma etapa crucial que permite extrair informações biológicas acerca do estado e tipo das células. Além disso, os agrupamentos encontrados são utilizados posteriormente na análise de diferenciação na expressão (DE) gênica das amostras sob diferentes condições biológicas, neste caso os grupos de pacientes controle (HC), com COVID-19 moderada (M) e severa (S). É importante considerar os agrupamentos obtidos na análise DE, uma vez que, os grupos capturam a heterogeneidade da expressão gênica das células e sob o ponto de vista da modelagem são considerados populações de células que devem ser incorporados de alguma forma na identificação de genes diferencialmente expressos entre os grupos de indivíduos do estudo (HC, M e S).

Neste trabalho, o agrupamento das células foi realizado utilizando o método *k-means* e considerando as 50 primeiras componentes principais obtidas a partir das expressões gênica log-normalizadas e corrigidas pelo efeito *batch* de indivíduo. Utilizar as componentes principais é motivado pelo fato de tirar vantagem da compressão dos dados e eliminar ruídos presentes nas expressão gênicas. O método *k-means* requer o conhecimento a priori do número de grupos (k) que serão considerados, assim para guiar a escolha de k, métricas estatísticas de separação dos grupos e a interpretação biológica de cada grupo foram investigadas. As métricas soma de quadrados total dentro de cada grupo (WSS) e a estatística GAP proposta por Tibshirani, Walther e Hastie (2002) utilizando 100 amostras *Bootstrap* foram os critérios adotados. Os resultados apresentadas na Figura 4.22 indicam que a medida que k aumenta tanto o WSS quanto a estatística GAP tendem estabilizar no mínimo (ou máximo) para valores próximos de $k = 20$. Analisando a estatística GAP verifica-se que o intervalo *Bootstrap* obtido para $k = 13$ engloba aproximadamente valores estimados para $k = 14, 15$ e 16.

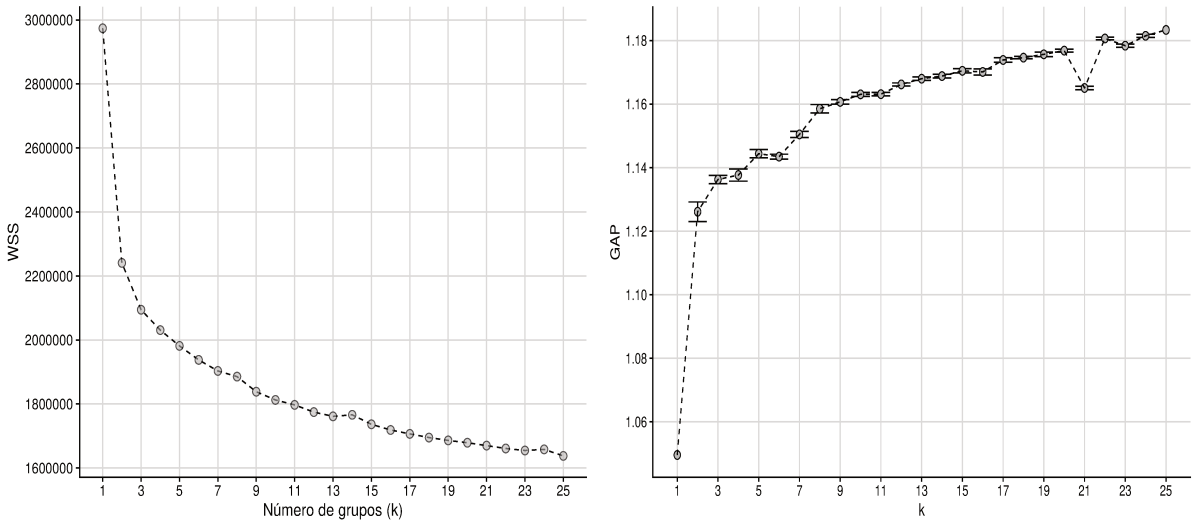


Figura 4.22 – Soma de quadrados dentro de cada grupo (WSS) e estatística GAP conforme o tamanho do grupo (k).

Nesse sentido, para auxiliar na escolha do k a anotação do tipo de célula predominante para cada grupo encontrado foi realizada. O conjunto de referência adotado foi o *Human Primary Cell Atlas* (HPCA), que consiste em 713 amostras com expressões gênicas log-normalizadas de 19.363 genes, sendo que para cada amostra foi atribuído um dos 37 principais tipos de células e 157 subtipos de células especificas (MABBOTT et al., 2013). A anotação foi realizada seguindo a proposta recentemente de Fu et al. (2019) e implementada na função `clustify` do **clustifyr** disponível no software R. Neste método, os grupos são anotados conforme a maior correlação de Spearman obtida entre as expressões gênicas do conjunto de referência (HPCA) e os dados da análise.

A Figura 4.23 ilustra os tipos de células predominantes em cada um dos grupos

encontrados. Por exemplo, para os agrupamentos obtidos com $k = 6$ indicam que os tipos de células predominantes são dois grupos com monócitos, dois grupos macrófagos, um grupo célula epitelial do brônquio e um grupo com célula B CD34. Essa mesma configuração dos tipos de células predominantes é observada para $k = 9$, com a adição de um grupo neutrófilo. Observa-se também um grupo de células T para $k = 20$. Percebe-se que quanto maior o número de k os grupos se tornam redundantes no que diz respeito ao tipo de célula predominante, em particular para $k = 17$ e 15 foram anotados 6 grupos como sendo macrófagos alveolar, ou seja, embora exista diferenças nas expressões gênicas desses grupos sob o ponto de vista da anotação do tipo de célula esses grupos são idênticos. Nesse sentido, para fixar um número de grupos, consideramos $k = 13$, uma vez que com essa quantidade de grupos é possível caracterizar seis principais grupos distintos de células, sendo eles células epiteliais, macrófagos, monócitos, neutrófilos e Pre-B_cell_CD34.

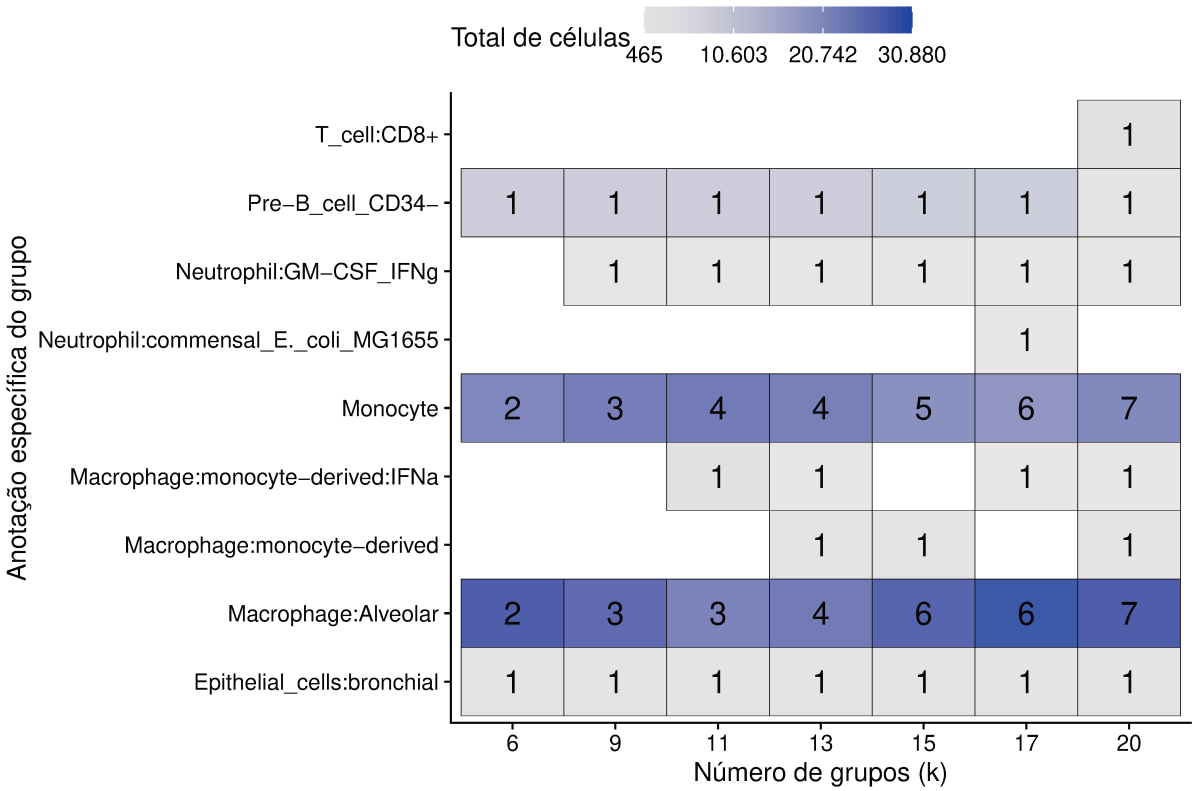


Figura 4.23 – *Heatmap* do total de células e tipo específico de célula anotado anotado para cada agrupamento. A escala colorida indica a quantidade de células presente e os números indicam a quantidade de grupos anotados para o determinado tipo de célula.

Uma característica desejada na análise de agrupamento é que os grupos encontrados sejam estáveis a pertubações dos dados, isto é, que pequenas mudanças nos dados não alterem as conclusões do estudo. Nesse sentido, para avaliar a estabilidade dos grupos o método *Bootstrap* não paramétrico reamostrando diretamente das células com 50

réplicas foi empregado. Uma métrica de estabilidade adotada para avaliar a estabilidade individual de cada grupo foi a frequência de co-atribuição para cada par de grupos foi avaliada. Valores altos na diagonal da Figura 4.24 indicam que o grupo correspondente é altamente consistente com o conjunto de dados original, enquanto que valores altos fora da diagonal fornecem indícios que par correspondente de grupos não são estáveis um em relação ao outro. Os resultados apresentados na Figura 4.24 apontam que os grupos 2, 5, 6, 7, 8, 9, 10, 11 e 12 são consistentes, enquanto que os demais apresentam uma pequena instabilidade que pode ser justificada pelo fato dos grupos 1, 3, 4 e 13 serem anotados como monócitos ou macrófagos derivados de monócitos (ver Tabela 4.7), ou seja, possuem semelhanças biológicas entre si. Uma métrica utilizada para avaliar a estabilidade geral do agrupamento foi o índice de Rand ajustado (ARI), calculado utilizando as 50 amostras *Bootstrap*. O ARI obtido foi de 0,7484, indicando que os grupos encontrados pelo método *k-means* aparentam ser robustos a ruídos amostrais.

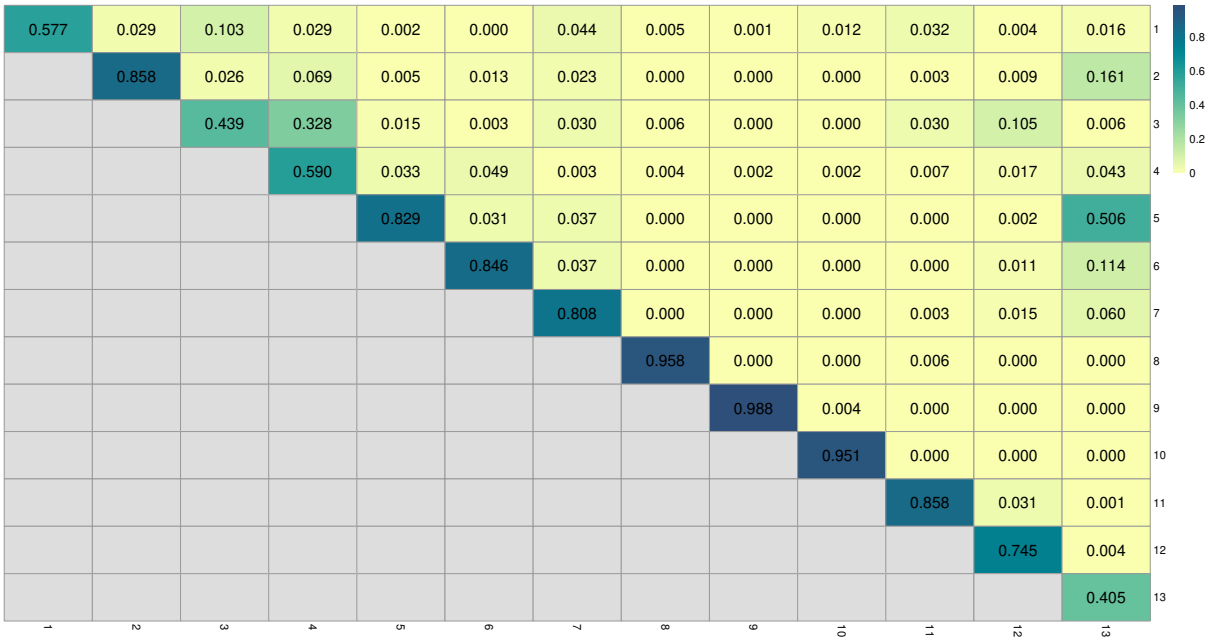


Figura 4.24 – *Heatmap* com a probabilidade de co-atribuição de cada par dos grupos originais obtido por 50 amostras *Bootstrap*. Cada quadrante representa a proporção de células atribuídas ao par de grupos.

Para melhor caracterização dos grupos encontrados a Figura 4.25 apresenta a proporção de células em cada grupo conforme indivíduo. Observa-se que os grupos 4 e 6 são predominantes de células de indivíduos com COVID-19 moderada, enquanto que os pacientes com COVID-19 severa estão em sua maioria presentes nos grupos 3, 5, 8, 9, 11, 12 e 13, em particular o grupo 8 apresenta 923 células de pacientes com COVID-19 severa e apenas 3 de indivíduos com COVID-19 moderada, como pode-se observar também na Tabela 4.6. Por outro lado, os grupos 1 e 3 são predominantes das células dos indivíduos controle.

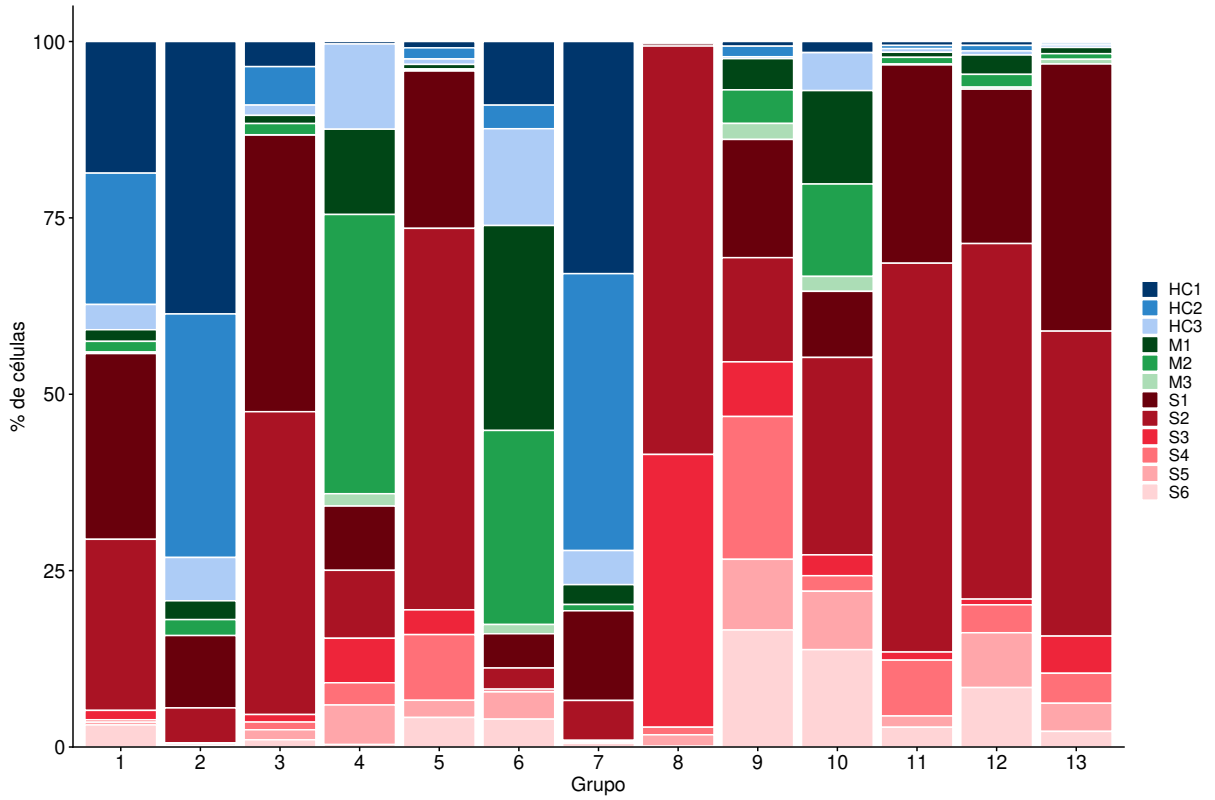


Figura 4.25 – Percentual de células por indivíduo conforme o grupo.

Tabela 4.6 – Frequência absoluta e percentual da quantidade de células em cada grupo conforme o status do indivíduo. O percentual entre parenteses refere-se a frequência relativa do número de células dentro de cada grupo.

Grupo	HC	M	S
1	2.008 (40,85%)	166 (3,38%)	2.741 (55,77%)
2	9.188 (79,28%)	571 (4,93%)	1.831 (15,80%)
3	732 (10,47%)	196 (2,80%)	6.064 (86,73%)
4	71 (12,43%)	305 (53,42%)	195 (34,15%)
5	76 (3,25%)	22 (0,94%)	2.243 (95,81%)
6	1.116 (26,08%)	2.476 (57,86%)	687 (16,06%)
7	5.112 (76,99%)	246 (3,70%)	1.282 (19,31%)
8	0 (0,00%)	3 (0,32%)	923 (99,68%)
9	44 (2,43%)	207 (11,45%)	1.557 (86,12%)
10	650 (6,96%)	2.658 (28,46%)	6.033 (64,59%)
11	85 (1,53%)	100 (1,79%)	5.387 (96,68%)
12	125 (1,92%)	316 (4,84%)	6.085 (93,24%)
13	19 (0,84%)	53 (2,35%)	2.186 (96,81%)

Como forma de visualização gráfica dos agrupamentos encontrados a represen-

tação em duas dimensões utilizando a redução de dimensionalidade UMAP das células de acordo com os grupos é apresentada na Figura 4.26. Pode-se notar que os grupos 9 e 10 se localizam num espaço mais afastados dos demais grupos. Tal fenômeno pode ser justificado pelo fato dos grupos 9 e 10 serem predominantes de células epiteliais e Pre-B_cell_CD34, respectivamente (ver Tabela 4.7). Isto é, as expressões gênicas em relação aos demais grupos é diferente e portanto refletida na representação UMAP, um subespaço com duas dimensão das expressões gênicas. Em relação aos demais grupos, percebe-se que há uma aproximação da representação UMAP dos genes, o que pode ser explicado pelo fato da maioria dos grupos, com exceção do grupo 8 serem caracterizados majoritariamente por células monócitos ou macrófagos (ver Tabela 4.7).

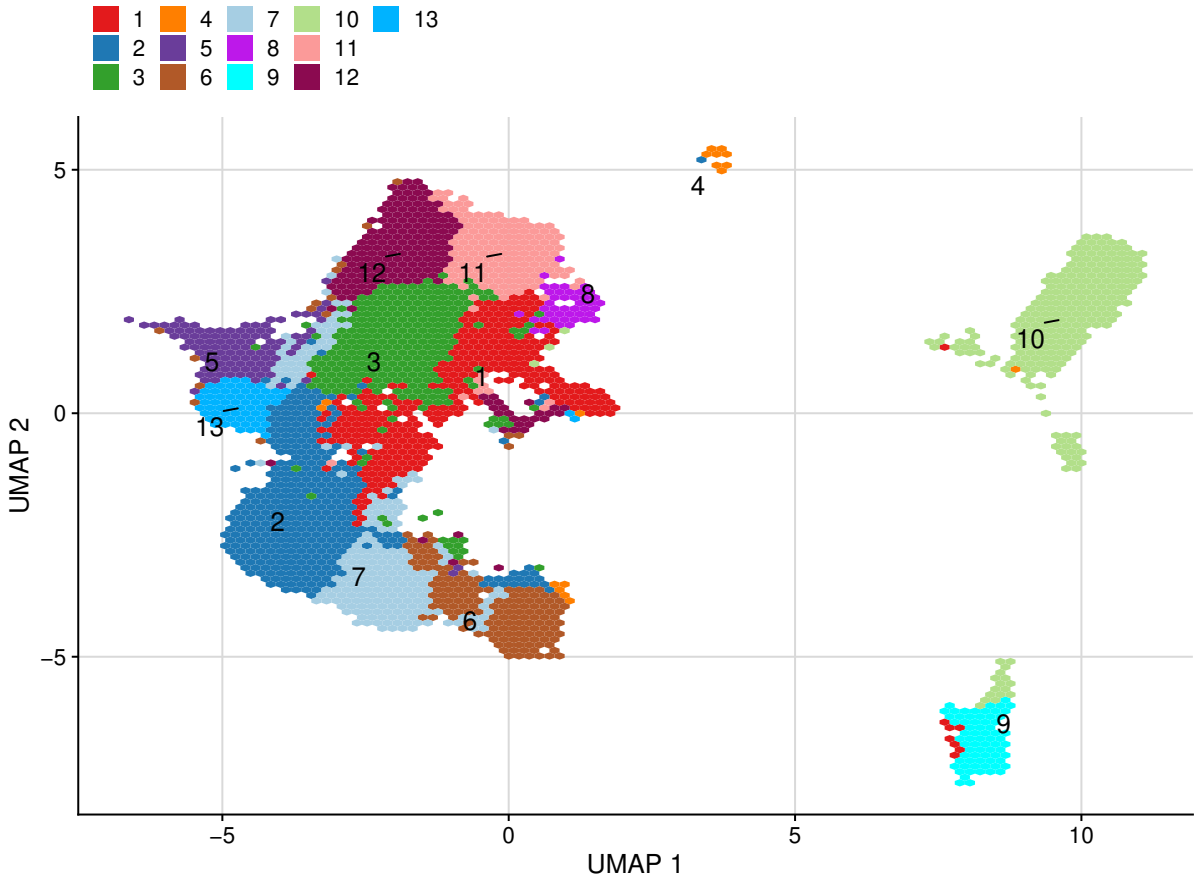


Figura 4.26 – Gráfico *hexbin* da representação UMAP das células conforme os grupos encontrados.

Tabela 4.7 – Anotação dos tipo de células predominantes para cada grupo ordenada conforme o total de células em cada grupo.

Grupo	Total de células	ρ	Principal	Específico
1	4.915	0,6479	Monocyte	Monocyte
2	11.590	0,6749	Macrophage	Macrophage:Alveolar
3	6.992	0,6868	Monocyte	Monocyte
4	571	0,6641	Macrophage	Macrophage:monocyte-derived:IFNa
5	2.341	0,6777	Macrophage	Macrophage:Alveolar
6	4.279	0,6853	Macrophage	Macrophage:Alveolar
7	6.640	0,6613	Macrophage	Macrophage:Alveolar
8	926	0,6480	Neutrophils	Neutrophil:GM-CSF_IFNg
9	1.808	0,6339	Epithelial_cells	Epithelial_cells:bronchial
10	9.341	0,6106	Pre-B_cell_CD34-	Pre-B_cell_CD34-
11	5.572	0,6985	Monocyte	Monocyte
12	6.526	0,7070	Monocyte	Monocyte
13	2.258	0,6918	Macrophage	Macrophage:monocyte-derived

De forma geral, os resultados encontrados da análise de caracterização das células BALF, em particular a análise de agrupamento realizada, corroboram com o que a ciência, especialmente a imunologia, descobriu sobre o ciclo biológico do vírus SARS-CoV-2 em seres humano até o presente momento. Por exemplo, a discussão realizada por Merad e Martin (2020) coincide com os grupos de células encontrados e analise de anotação dos tipos de células predominantes em cada grupo.

Especificamente, as células epiteliais possuem função de revestimento de tecidos e órgãos, tais como o pulmão, e conhecidamente são capazes de expressar a Enzima Conversora de Angiotensina (ECA2), que é uma das proteínas-chave para a ligação do SARS-Cov-2 à célula e, portanto, para que a infecção viral ocorra. De forma condizente, observa-se que o Grupo 9, é caracterizado por células epiteliais do brônquio (ver Tabela 4.7), e além disso é quase que exclusivamente oriundo de células de indivíduos com COVID-19, conforme apresentado na Figura 4.25 e Tabela 4.6.

O tipo de célula mais frequente entre os grupos encontrados foi o macrófago, particularmente anotado nos grupos 2, 4, 5, 6, 7 e 13. Os macrófagos são células do sistema imune inato que possuem ação fagocíticas, ou seja, têm como função principal o engolfamento e destruição de partículas, tais como patógenos e debris celulares. Essas células também agem como agentes sinalizadores da resposta infamatória, isto é, ao realizarem a fagocitose, os macrófagos são capazes de realizar o recrutamento celular, através da liberação de citocina e quimiocinas para os sítios da infecção. Um tipo de

célula capaz de responder a esse recrutamento são os neutrófilos, encontrados no Grupo 9, que são células fagocíticas circulantes e são atraídas até o sitio da infecção para combate do patógeno. Este comportamento biológico descrito está coerente com os resultados encontrados, e é parte do sistema imune inato, componente do sistema imune que é responsável pelo reconhecimento e combate de patógenos através da identificações de padrões moleculares amplamente difundidos na natureza, como componentes da membrana bacteriana, RNA dupla fita, entre outros.

De forma complementar, existe a resposta imune adaptativa, cuja principal propriedade é reconhecer características específicas de um microrganismo, criando uma resposta direcionada ao combate daquele patógeno e, finalmente, gerando uma memória imunológica. No contexto de uma infecção por SARS-CoV-2, os macrófagos, responsáveis pela fagocitose do material viral, realizam a apresentação do antígeno viral através do complexo principal de histocompatibilidade (MHC). Essa apresentação do antígeno, ou seja, molécula capaz de deflagrar a produção de anticorpos, é feita para as células T.

As células T ou linfócitos T são responsáveis por orquestrar a defesa do organismo contra patógenos, de forma a torná-la cada vez mais específica devido ao reconhecimento dos antígenos. Para que esse reconhecimento aconteça, os linfócitos T expressam em sua membrana receptores específicos dessa linhagem celular (TCR). Uma vez que o complexo MHC e TCR entram em contato, a célula T passa a reconhecer o antígeno que foi apresentado pelo fagócito, e é capaz de disparar a resposta imune específica àquele antígeno. Essa resposta inclui a proliferação dos linfócitos T e encaminhamento até o sítio da infecção, no caso do SARS-CoV-2, o pulmão. No sitio da infecção, a célula T media a resposta das demais células do sistema imune, através da liberação de citocinas e quimiocinas, o que possibilita o recrutamento dos monócitos circulantes (grupos 1, 3, 11 e 12) que quando chegam no sítio da infecção, se diferenciam em macrófagos (ver a coluna Específico da Tabela 4.7) e assumem função efetora para eliminação do patógeno.

Outra função crucial da célula T é sinalizar e induzir a formação de anticorpos nas células B, predominantes no Grupo 10. Anticorpos são proteínas plasmáticas liberadas pela célula B capazes de se ligar ao antígeno possibilitando a sinalização da presença do antígeno no organismo e, assim, regular a atividade das demais células do sistema imune, conferindo especificidade à resposta.

4.2.3 Detecção de marcadores gênicos para cada agrupamento

A análise de agrupamento combinada com a anotação dos grupos fornecem uma caracterização geral do comportamento biológico das células presentes em cada grupo. Para complementar e melhor caracterizar cada um grupos de células encontrados é usual procurar genes marcadores, isto é, genes que são diferencialmente expressos conforme entre os agrupamentos. A identificação dos genes que realizam a separação entre os grupos

permite atribuir maior significado biológico a cada grupo com base na anotação funcional dos genes.

É importante ressaltar que as estratégias para encontrar genes marcadores entre os grupos obtidos na análise de agrupamento são estatisticamente falhas até certo ponto, uma vez que a análise é realizada nos mesmos dados utilizados para encontrar os grupos, ou seja, a hipótese de interesse – existem diferenças entre os grupos? – é formulada a partir dos dados onde os grupos foram encontrados e como os métodos de agrupamento naturalmente fornecem grupos diferentes é mais provável obter resultados positivos. Esse fenômeno é conhecido na literatura como *data dredging* (SMITH; EBRAHIM, 2002). Todavia, como discutido na Seção 3.2, para detecção de genes marcadores os valores-p dos testes são utilizados para ranquear os genes e portanto o efeito do *data dredging* se torna amplamente inofensivo. Em contrapartida, se os resultados dos testes forem utilizados para interpretar diferenças estatisticamente significativas entre os grupos, tais resultados serão equivocados, pois o conceito de significância estatística para diferenças entre grupos não é aplicável se os mesmos e suas interpretações não forem reproduzíveis de forma estável em experimentos replicados.

Para identificar genes marcadores, testes de comparações múltiplas foram realizados seguindo o estado da arte discutido no Capítulo 4. Em particular, três estratégias foram adotadas afim de comparar os resultados, sendo elas: (i) teste t de Welch blocando por indivíduo; (ii) teste de Wilcoxon blocando por indivíduo e (iii) modelo linear com efeito fixo de indivíduo. Essas abordagens estão consolidadas, respectivamente, nas funções `pairwiseTTest(sce, block)`, `pairwiseWilcox(sce, block)` e `pairwiseTTest(sce, design)`, do pacote **scran**. Ressalta-se que, devido a presença de vários indivíduos dentro de um mesmo grupo, conforme ilustram a Figura 4.25 e Tabela 4.6, devem ser incorporados na modelagem, pois podem interferir na detecção dos genes marcadores, mais precisamente inflando a variância dentro de cada grupo ou também distorcendo as mudanças *log-fold* se a composição do grupo variar entre os indivíduos.

Os resultados da Figura 4.27 indicam o total de genes marcadores considerando o FDR menor que 5% para cada grupo de acordo como a abordagem adotada. Observa-se que o modelo linear identificou menos genes marcadores em todos os grupos do que os testes t de Welch e Wilcoxon, isso pode ser justificado pelo fato da variância estimada nas comparações múltiplas entre grupos ser igual entre os genes, isto é, a variância do resíduo. Por outro lado, a abordagem utilizando o teste t de Welch relaxa a suposição de variância igual entre os grupos e consequentemente apresenta maior número de gene marcadores. O teste de Wilcoxon detectou mais genes marcadores nos grupos 4 e 9 do que os demais métodos.

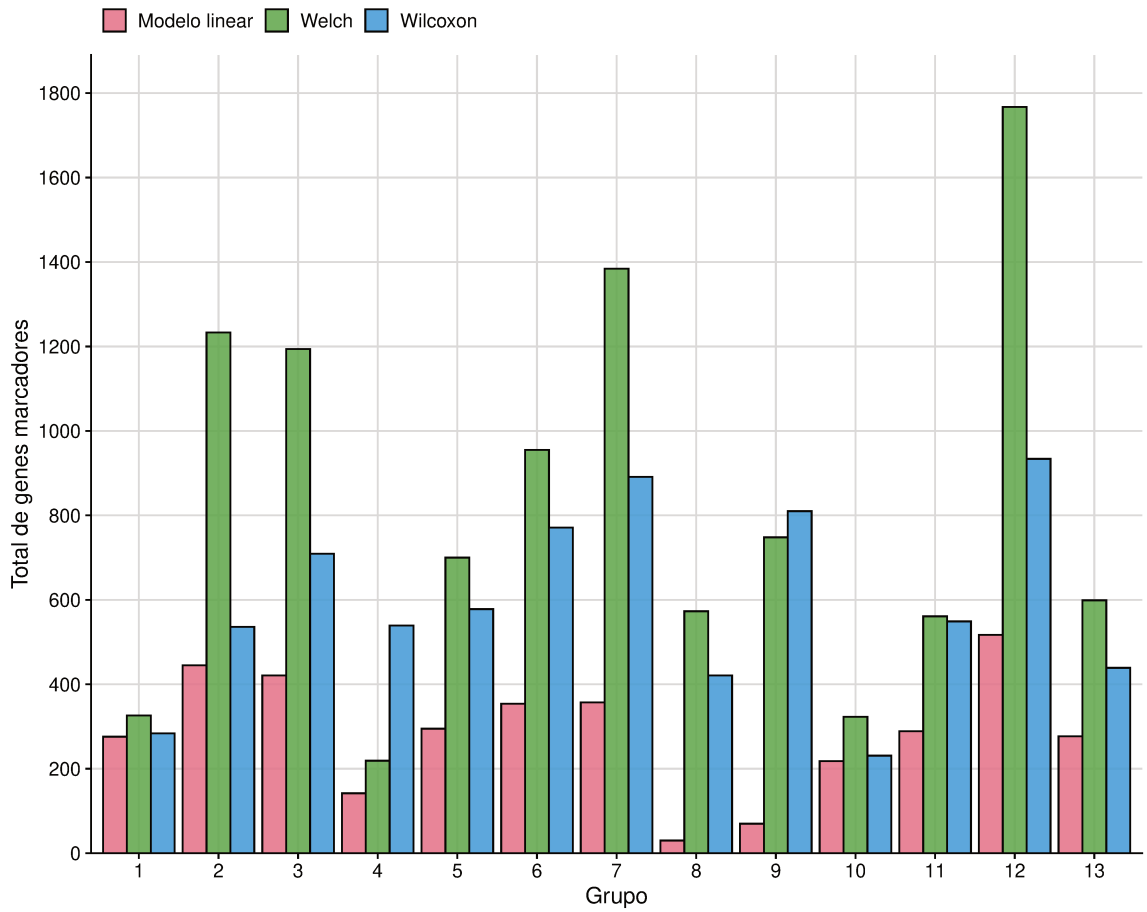


Figura 4.27 – Total de genes marcadores considerando FDR menor que 5% em cada grupo conforme método.

A Figura 4.28 apresenta a proporção de genes marcadores compartilhados entre os métodos. Os top genes marcadores são definidos como aqueles que apresentaram os menores 20, 50 ou 100 FDR das comparações múltiplas entre os grupos. Observa-se que a abordagem Welch e Modelo linear apresentam alta concordância dos para alguns grupos, particularmente 1, 2, 5, 10, 11 e 13 a proporção de genes marcadores entre os dois métodos é superior a 50%. Por outro lado, nos demais grupos observa-se que a concordância entre os métodos é menor que 40%, independente da quantidade de top genes (20, 50 ou 100). O teste de Wilcoxon, em geral, detecta no máximo 30% dos top 20 genes iguais ao modelo linear ou teste t de Welch, particularmente para os grupos 1, 10 e 11.

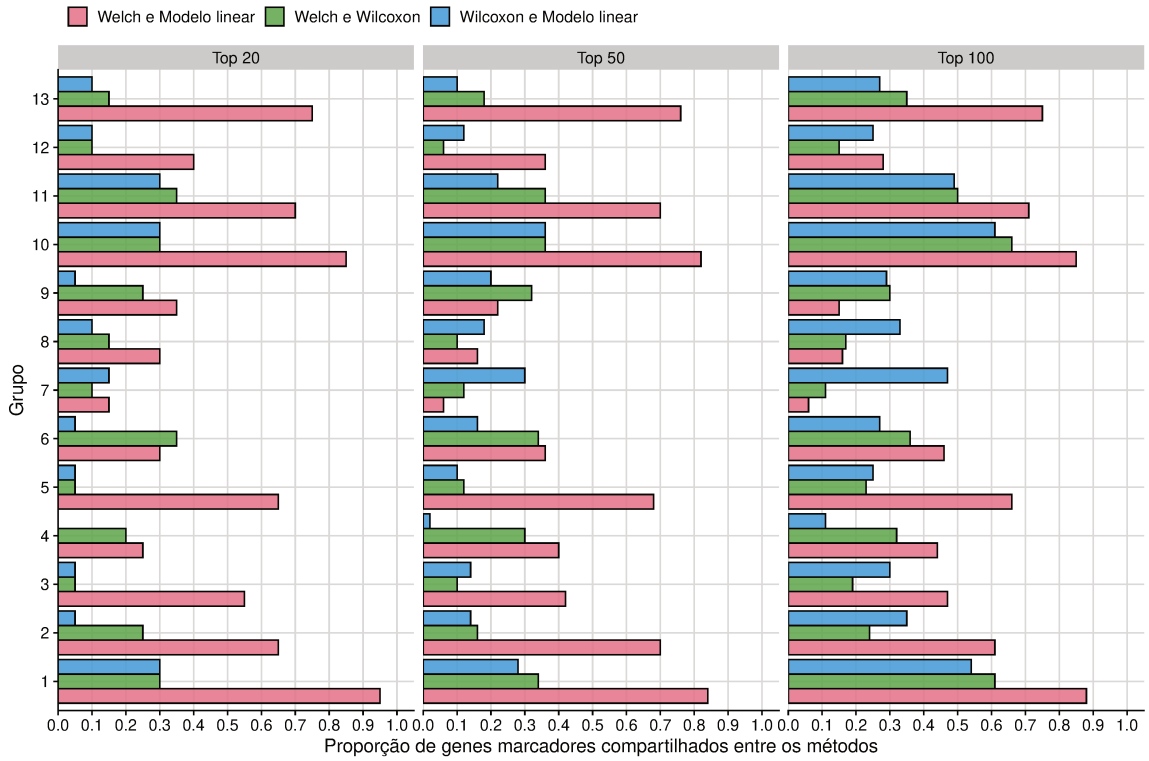


Figura 4.28 – Proporção dos top genes marcadores compartilhados entre métodos. Os top genes foram definidos como aqueles com os menores 20, 50 ou 100 FDR das comparações múltiplas.

A comparação entres os métodos enfatiza que a detecção de genes marcadores é suscetível ao método adotado, e portanto, a escolha adequada da abordagem, conforme a estrutura experimental dos dados bem como as características dos testes é fundamental para encontrar os genes candidatos a marcadores genéticos.

O teste de Wilcoxon-Mann-Whitney (WMW) possui algumas vantagens em relação ao teste t de Welch e o modelo linear. Por ser um teste não paramétrico é menos restringente em suas suposições, além disso, avalia diretamente a separação entre as distribuições da expressão gênica de diferentes grupos. A estatística do teste WMW é proporcional a área sob a curva ROC (AUC), isto é, a probabilidade de concordância de uma célula aleatório em um grupo ter uma expressão mais alta do que uma outra célula aleatório em outro grupo. Nas comparações múltiplas valores de AUC próximos de 0 ou 1 indicam que o gene separa perfeitamente os dois grupos. Dessa forma, o teste WMW aborda diretamente a propriedade mais interessante de um candidato a gene marcado, enquanto que o teste t de Welch e o modelo linear apenas o fazem indiretamente por meio das diferenças nas médias e variâncias entre os grupos (AMEZQUITA et al., 2020). As Figuras 4.29 e 4.30 apresentam a AUC dos top 5 genes para cada grupo.

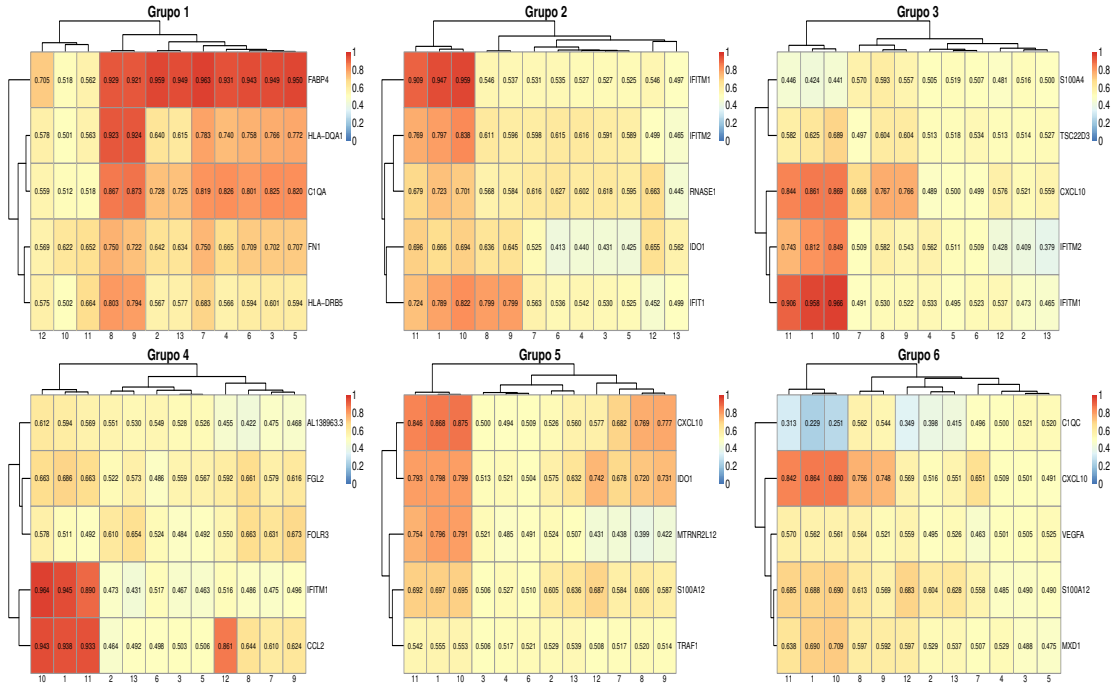


Figura 4.29 – *Heatmap* com as AUC dos top 5 genes ranqueados pelo valores-p dos testes de comparações múltiplas entre os grupos – Grupos 1 a 6.

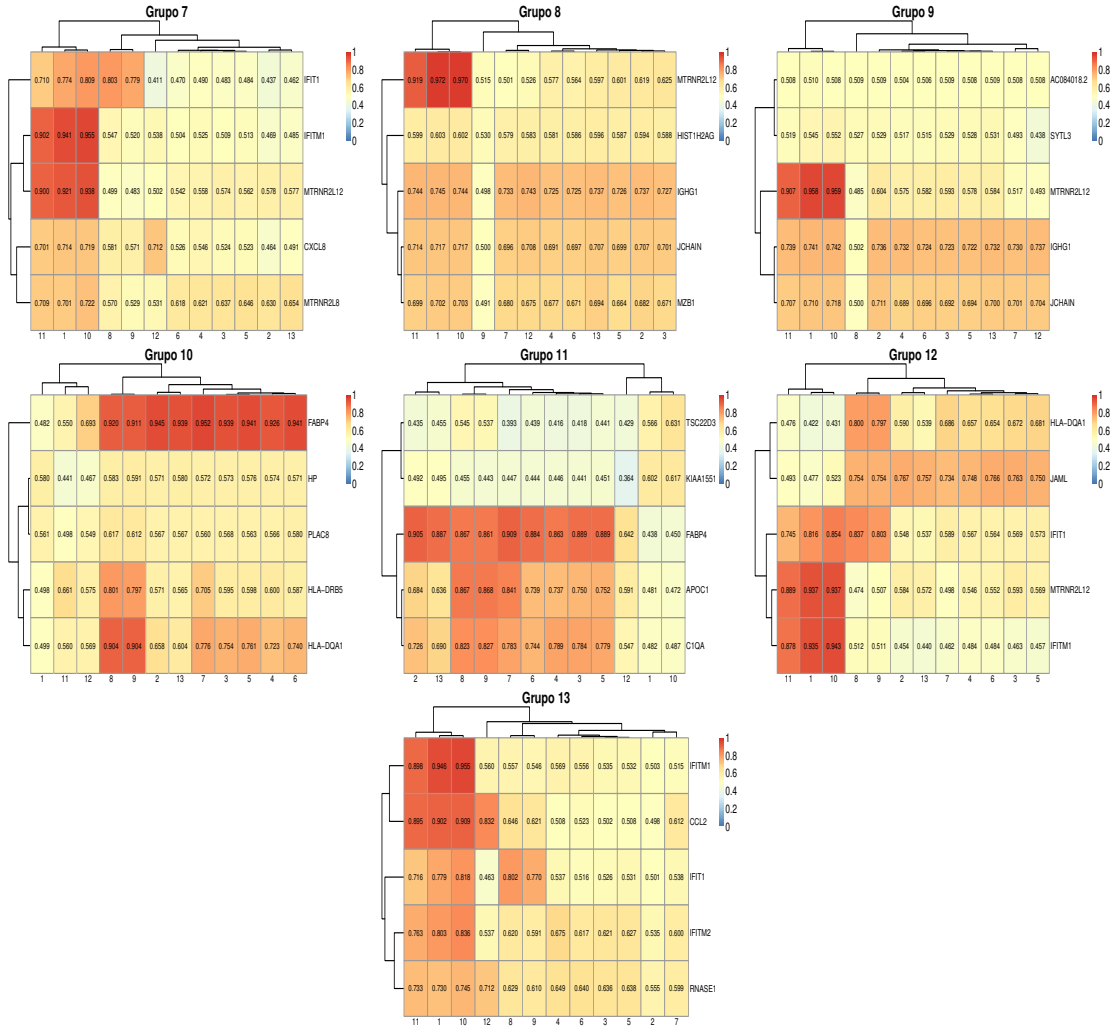


Figura 4.30 – *Heatmap* com as AUC dos top 5 genes ranqueados pelo valores-p dos testes de comparações múltiplas entre os grupos – Grupos 7 a 13.

Na Figura 4.31 a expressão média dos genes marcadores detectados pelo teste de Wilcoxon é apresentada em cada grupo. Observa-se, por exemplo, que os grupos 1, 10 e 11, apresentam expressões médias similares para a maioria dos genes marcadores.

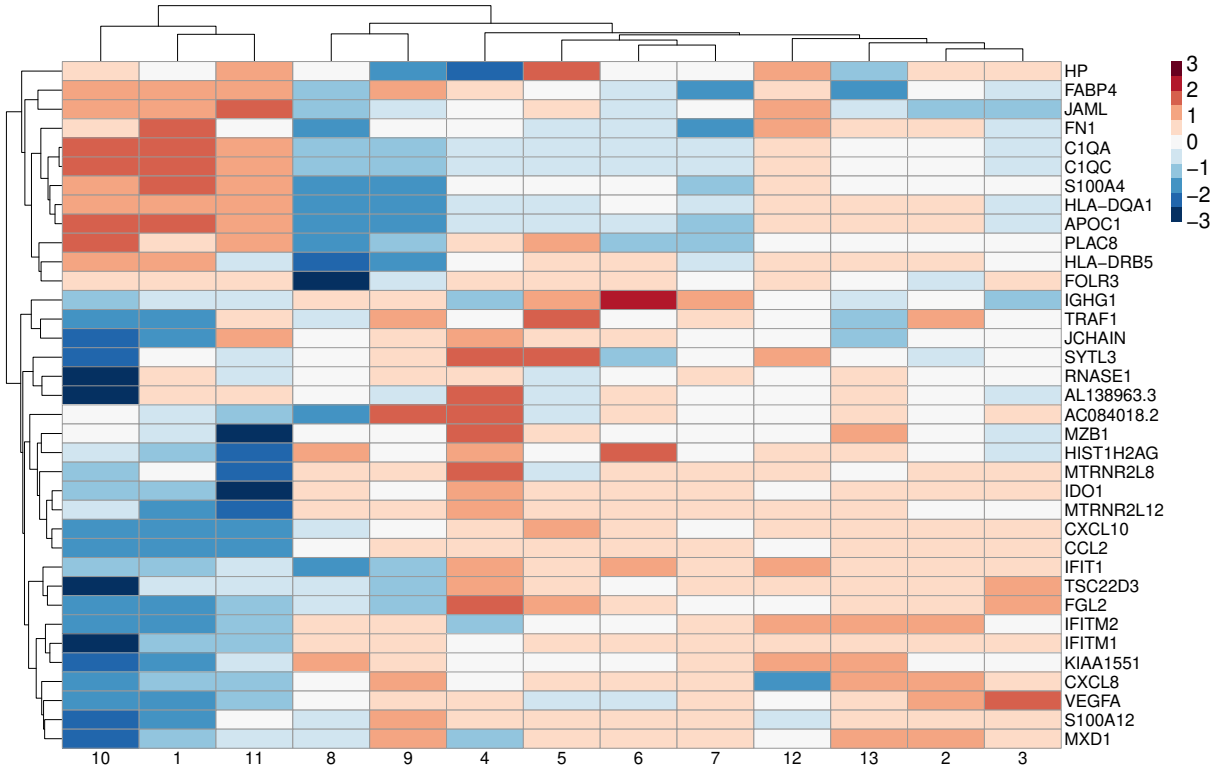


Figura 4.31 – *Heatmap* da expressão média normalizada e padronizada por gene dos genes marcadores detectados pelo teste de Wilcoxon.

4.2.4 Análise da expressão gênica conforme a condição dos indivíduos

Após caracterização dos grupos de células, a partir da análise de agrupamento e detecção de genes marcadores, têm-se o objetivo em investigar e identificar genes expressos diferencialmente conforme os grupos dos indivíduo, isto é, entre os indivíduos controle (HC), com COVID-19 moderada (M) e severa (S). Obviamente, os grupos de células encontrados a partir dos dados devem ser considerados na análise, uma vez que, no nível celular é observado uma heterogeneidade na expressão gênica dentro de cada grupo de indivíduo, a qual espera-se controlar a partir dos grupos encontrados e anotados (ver Tabela 4.7). Dessa forma, a análise da diferenciação na expressão (DE) gênica foi conduzia incorporando na modelagem os 5 tipos principais de grupos celulares encontrados, conforme reportado na Tabela 4.7 e Figura 4.23. Ressalta-se que o Grupo 8, anotado como células do tipo neutrófilos, não foi considerado na analise uma vez que das 925 células atribuídas a este grupo 923 são de indivíduos com COVID-19 severa. A Tabela 4.8 apresenta a quantidade de células em cada grupo celular conforme os grupos dos indivíduo. Os macrófagos são os tipos de células com maior predominância, sendo 56% provenientes de indivíduos controle, 14% de pacientes com COVID-19 moderada e 30% com COVID-19 severa.

Tabela 4.8 – Distribuição da quantidade de células em cada grupo celular conforme os grupos dos indivíduo. O percentual entre parênteses refere-se a frequência relativa do número de células dentro de cada grupo.

Tipo de célula	HC	M	S
Epithelial_cells	44 (2,43%)	207 (11,45%)	1.557 (86,12%)
Macrophage	15.582 (56,30%)	3.673 (13,27%)	8.424 (30,43%)
Monocyte	2.950 (12,29%)	778 (3,24%)	20.277 (84,47%)
Pre-B_cell_CD34-	650 (6,96%)	2.658 (28,46%)	6.033 (64,59%)

A presença de células provenientes de diferentes indivíduos introduz na análise DE um fator de variabilidade extra inerente à origem da célula, denominado na literatura estatística de efeito aleatório. Ignorar a origem da célula implica que a expressão gênica de cada célula são réplicas independentes. Essa abordagem não apresenta o rigor estatístico adequado, uma vez que a variabilidade devido ao indivíduo não é propriamente modelada. Neste sentido, a modelagem da expressão gênica foi realizada abordando duas estratégias: (i) modelo misto com efeito aleatório de individuo e modelando a expressão gênica $\log_2$-normalizada pelo método *deconvolution*; (ii) modelo *edgeR* proposto por Robinson, McCarthy e Smyth (2009) modelando as contagens obtidas por amostras pseudo-*bulk*, assumindo uma distribuição Binomial Negativa.

Seja y_{gijlk} a expressão gênica $\log_2$-normalizada do gene g na célula i do indivíduo j com fenótipo l no grupo celular k, então o modelo misto ajustado para cada gene $g = 1, \ldots, G$ de forma independente, pode ser definido utilizando a notação de modelos multiníveis da seguinte forma

$$y_{gijlk} = \alpha_{gl} + \gamma_{gk} + (\alpha\gamma)_{glk} + b_{gj} + \varepsilon_{gilk} \tag{4.2}$$

em que $i = 1, 2, \ldots, n_j$ células; $j = 1, 2, \ldots, 12$ indivíduos; $k = \text{HC}, \text{M}, \text{S}$ fenótipos; $l = 1, 2, 3, 4$ grupos celulares apresentados na Tabela 4.8; $b_{gj} \sim N(0, \tau_g^2)$ é a componente aleatório devido ao indivíduo para cada gene g e $\varepsilon_{gilk} \sim N(0, \sigma_g^2)$ representa a componente aleatória do modelo.

O modelo definido em (4.2) apresenta parâmetros para os efeitos fixos de fenótipo (α_{gl}), grupo celular (γ_{gk}) e a interação entre fenótipo e grupo celular $(\alpha\gamma)_{glk}$. O termo de interação foi incluído no modelo, pois verifica-se que as expressões gênicas entre os grupos de indivíduos diferem conforme os tipos das células. Sob o ponto de vista prático o modelo (4.2) é reescrito na notação matricial utilizando variáveis *dummies*, da seguinte forma:

$$\boldsymbol{y}_g = \mathbf{X}\boldsymbol{\beta}_g + \mathbf{Z}\boldsymbol{u}_g + \boldsymbol{\varepsilon}_g \tag{4.3}$$

em que a matriz de experimento $\mathbf{X}$ tem dimensão 62833×12, com o vetor de parâmetros dado por $\boldsymbol{\beta}_g = (\beta_{g1}, \beta_{g2}, \ldots, \beta_{g12})^\top$, $\mathbf{Z}$ é a matriz de experimento do efeito aleatório e $\boldsymbol{u}_g$ é o vetor de parâmetros do efeito aleatório. O grupo de células epiteliais foi utilizado como referência e o modelo foi parametrizado sem intercepto.

Para melhor entendimento dos parâmetros modelo a Tabela 4.9 apresenta as interpretações dos parâmetros de acordo com efeito médio marginal para cada grupo e tipo de célula. Por exemplo, o coeficiente β_{g1} representa a média da expressão gênica das células epiteliais do grupo controle (HC).

Tabela 4.9 – Interpretação dos parâmetros de efeito fixo do modelo (4.2).

Grupo	Tipo de célula	Efeito médio marginal
HC	Epitelial	β_{g1}
M		β_{g2}
S		β_{g3}
HC	Macrófago	$\beta_{g1} + \beta_{g4}$
M		$\beta_{g2} + \beta_{g4} + \beta_{g7}$
S		$\beta_{g3} + \beta_{g4} + \beta_{g8}$
HC	Monócito	$\beta_{g1} + \beta_{g5}$
M		$\beta_{g2} + \beta_{g5} + \beta_{g9}$
S		$\beta_{g3} + \beta_{g5} + \beta_{g10}$
HC	Pre-B_cell_CD34-	$\beta_{g1} + \beta_{g6}$
M		$\beta_{g2} + \beta_{g6} + \beta_{g11}$
S		$\beta_{g3} + \beta_{g6} + \beta_{g12}$

A partir das interpretações apresentadas na Tabela 4.9 das médias marginais do modelo (4.2), utiliza-se contrastes para testar diferenças nas expressões gênicas de acordo com os grupos HC, M e S dentro de cada grupo celular. Sem perdas de generalidade, a hipótese formulada é definida por:

$$\mathcal{H}_0 : \boldsymbol{c}^\top \boldsymbol{\beta}_g = \mathbf{0} \tag{4.4}$$

em que $\boldsymbol{c}^\top$ é o vetor de contraste. Por exemplo, para testar se há diferença nas expressões gênicas das células dos indivíduos HC versus M dentro do grupo celular Epitelial têm-se que $\boldsymbol{c} = (1, -1, 0, 0, \ldots, 0)$. Para avaliar a diferença entre indivíduos HC e M dentro dos Macrófagos o vetor de contraste é definido por $\boldsymbol{c} = (1, -1, 0, 0, 0, 0, -1, 0, \ldots, 0)$. Para testar a hipótese (4.4) a estatística t, discutida no Capítulo 4, com os graus de liberdade baseados na aproximação do método do momento de Satterthwaite proposto por Giesbrecht e Burns (1985) foi utilizada.

Uma etapa usual na análise DE em dados de scRNA-seq é filtrar genes que apresentam uma abundância de contagens iguais a zero. Os ajustes dos modelos mistos considerou apenas os genes que tiveram menos que 90% das células com contagem observada igual a zero. Dessa forma, dos 2209 genes inicias 515 foram investigados na análise DE.

Para o modelo edgeR uma amostra pseudo-*bulk* foi gerada somando as contagens das células nos níveis dos indivíduos dentro dos 13 grupos encontrados. A Tabela 4.10 apresenta o número de amostras pseudo-*bulk* para cada tipo de célula conforme os grupos de indivíduos. Somar as contagens remove o efeito devido ao indivíduo e os métodos já estabelecidos para análise de *bulk* RNA-seq podem ser utilizados, além disso, estudos de simulação realizados por Lun e Marioni (2017) e apresentados neste trabalho evidenciam que não há perda no controle do erro do Tipo I, bem como do poder dos testes.

Tabela 4.10 – Distribuição da quantidade de observações pseudo-*bulk* em cada tipo de célula conforme o fenótipo do indivíduo. O percentual entre parênteses refere-se a frequência relativa dentro de cada grupo.

Tipo de célula	HC	M	S
Epithelial_cells	3 (25,00%)	3 (25,00%)	6 (50,00%)
Macrophage	17 (24,64%)	16 (23,19%)	36 (52,17%)
Monocyte	12 (25,00%)	12 (25,00%)	24 (50,00%)
Pre-B_cell_CD34-	3 (25,00%)	3 (25,00%)	6 (50,00%)

Para incorporar a heterogeneidade do tipo de célula na abordagem pseudo-*bulk* um modelo foi ajustado para cada tipo de célula e testes da razão de verossimilhança foram realizados utilizando os devidos contrastes para avaliar a diferença entre os grupos. Os modelos foram ajustados utilizando o pacote **edgeR** (ROBINSON; MCCARTHY; SMYTH, 2009), em particular, as função `estimateDisp` para estimação dos parâmetros de dispersão da Binomial Negativa para cada gene, `glmFit` para o ajuste do modelo e `glmLRT` para os testes da razão de verossimilhanças. Detalhes teóricos sobre os procedimentos adotados foram discutidos no Capítulo 3.

Na análise de bulk RNA-seq algumas etapas de pré-processamento são realizadas previamente ao ajustes dos modelos. Nesse sentido, seguindo as propostas descritas em Chen, Lun e Smyth (2016) dois pré-processamentos foram realizados, (i) remoção de genes que apresentam contagem total inferior 10 em pelo menos alguma amostra pseudo-*bulk* e (ii) cálculo do fator de normalização utilizado no modelo como *offset* para cada amostra pseudo-*bulk* utilizando a estratégia *trimmed mean of M-values* proposta por Robinson e Oshlack (2010). Importante ressaltar que as estratégias de filtragem inicial dos genes foram diferentes para o modelo misto e edgeR.

Para cada modelo os valores-p dos testes de hipóteses foram corrigidos utilizando a correção de BH proposta por Benjamini e Hochberg (1995) para controlar o *false discovery rate* (FDR). O uso do valor-p corrigido é uma prática comum na análise de estudos genômicos, uma vez que, múltiplos testes são simultaneamente conduzidos para identificação de genes diferencialmente expressos, e portanto, para controlar falsas descobertas, em particular, genes falsos positivos, isto é, presença de diferenças quando não há, métodos para correção do valor-p são empregados. Assumindo que as suposições do teste são válidas, o valor-p corrigido pelo método de BH garante que se todos os genes com valor-p corrigido abaixo de um limiar são selecionados como diferencialmente expressos, então a proporção de descobertas falsas no grupo selecionado é controlada para ser inferior ao valor de limiar estabelecido. Para simplificação da linguagem, ao longo do texto o valor-p pelo método de BH corrigido será referido como FDR.

As Figuras 4.32 e 4.33 comparam os genes diferencialmente expressos identificados pelo modelo linear misto (MLM) e edgeR. Considerando nível de 5% para o FDR ambos os modelos não identificaram nenhum gene diferencialmente expressos em nenhuma comparação das células Epiteliais. Nos Macrófagos, apenas o MLM identificou genes diferencialmente expressos considerando o FDR de 5%, sendo 299 entre HC e M, 87 entre HC e S, e 255 entre M e S. A Figura 4.33 ilustra a proporção de genes DE compartilhados entre os métodos. Considerando os 20 primeiros genes com menor FDR os métodos apresentam uma concordância de cerca de 5% para todas as comparações nas células Epiteliais. Já nos Macrófagos considerando os 50 primeiros genes os métodos apresentam uma concordância de 10% nas comparações HC versus M e HC versus S. As menores concordâncias entre os métodos é observada nas células Monócitos, enquanto que as maiores nas células Pre-B-CD34-, neste último, atingindo 40% considerando os 100 primeiros genes na comparação HC versus S. Em geral, observa-se uma baixa ou em alguns casos nenhuma concordância entre os genes DE identificados pelos dois modelos. A diferença entre a quantidade bem como os genes diferencialmente expressos detectados pelos dois modelos pode ser explicada pela natureza da abordagem, ou seja, o modelo edgeR avalia diferenças nos genes a partir das amostras pseudo-*bulk*, enquanto que o modelo linear misto apresenta maior sensibilidade, pois as diferenças nas expressões gênicas são detectadas no nível da célula.

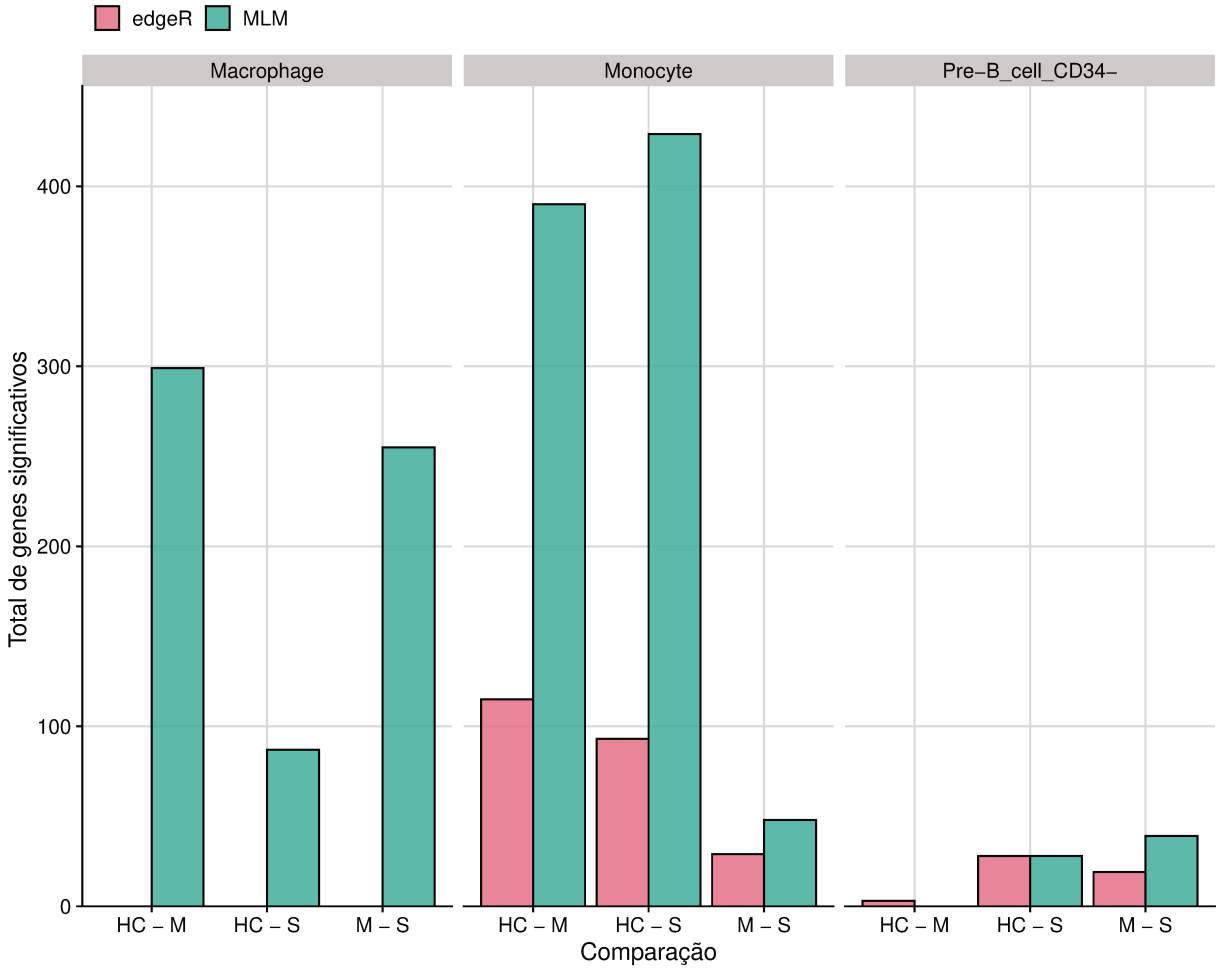


Figura 4.32 – Total de genes significativos considerando FDR menor que 5% em cada contraste conforme o modelo.

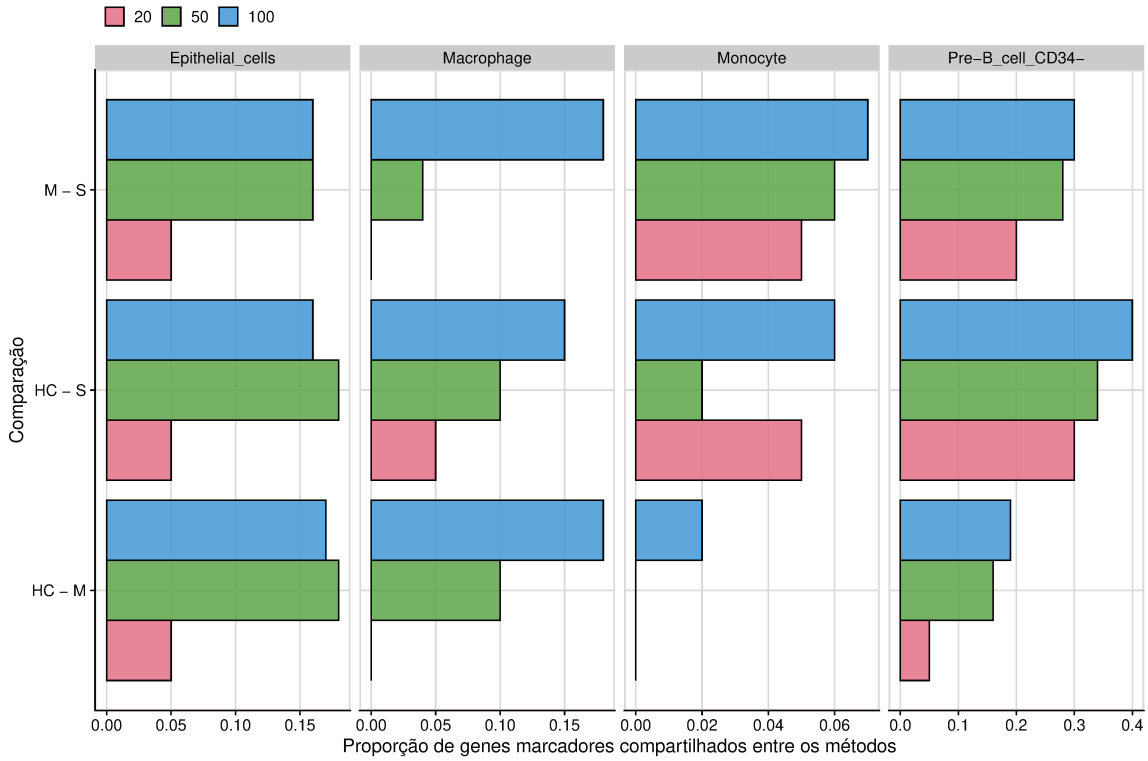


Figura 4.33 – Proporção dos top genes compartilhados entre modelos misto e edgeR. Os top genes foram definidos como aqueles com os menores 20, 50 ou 100 FDR das comparações múltiplas.

As Tabelas 4.11 e 4.12 apresentam os genes que obtiveram maior significância estatística por tipo de célula e comparação para os modelos misto e edgeR, respectivamente. Por exemplo, segundo o modelo linear misto, nas células do tipo Monócito, o gene CRIP1 apresenta uma regulação no indivíduos controle comparado com o grupo moderado, em particular, observa-se que este gene tem uma expressão aumentada de $2^{0,8487} = 1,8008$ vezes nas células dos indivíduos controle comparado com o grupo moderado. De forma geral, as estimativas do log-*fold change* (logFC) do modelo linear misto apresentaram uma magnitude baixa, nesse sentido, os resultados encontrados devem ser discutidos com pesquisadores da área para validação da significância biológica.

Por outro lado, as estimativas do logFC obtidas pelo modelo edgeR apresentam valores próximos e até superiores aos usuais limiares adotados pelo comunidade científica de ± 2 ou 4 (MCCARTHY; SMYTH, 2009). Por exemplo, segundo o modelo edgeR o gene S100P nas células do tipo Monócito apresentou um logFC de -4,45 na comparação do grupo controle com o moderado, isso implica num baixo nível de regulação de $-2^{4,45} = -21,85$ no grupo controle versus moderado. Interessante mencionar que, embora as estimativas do logFC do modelo edgeR apresentem uma intensidade maior, a significância estatística considerando o FDR não é encontrada para todos os genes na células Epiteliais e Macrófagos.

Tabela 4.11 – Resumo do gene com o menor valor-p obtidos pelo modelo linear misto conforme o tipo de célula e comparação entre os grupos.

Gene	Tipo de célula	Comparação	logFC	valor-p	FDR
FKBP11	Epithelial_cells	HC - M	0,2718	2,53E-03	9,97E-01
SLC7A5	Epithelial_cells	HC - S	0,2540	1,69E-04	2,62E-01
FKBP11	Epithelial_cells	M - S	-0,1381	9,98E-04	7,71E-01
SMIM25	Macrophage	HC - M	-0,3106	4,26E-21	6,58E-18
RHOB	Macrophage	HC - S	0,0755	1,55E-05	8,04E-05
FTH1	Macrophage	M - S	0,4407	9,94E-17	2,19E-14
CRIP1	Monocyte	HC - M	0,8487	1,13E-35	1,95E-33
CSTA	Monocyte	HC - S	0,4746	5,88E-148	9,09E-145
HLA-DRB5	Monocyte	M - S	0,4286	2,53E-08	9,40E-08
CTNNB1	Pre-B_cell_CD34-	HC - M	0,1015	1,49E-02	1,87E-01
S100A13	Pre-B_cell_CD34-	HC - S	0,2394	3,62E-11	3,34E-08
CRIP1	Pre-B_cell_CD34-	M - S	0,3361	4,32E-11	3,34E-08

Tabela 4.12 – Resumo dos genes com o menor valor-p obtidos pelo modelo edgeR conforme o tipo de célula e comparação entre os grupos.

Gene	Tipo de célula	Comparação	logFC	valor-p	FDR
NUSAP1	Epithelial_cells	HC - M	1,6336	3,46E-04	0,4128
SCGB1A1	Epithelial_cells	HC - S	-3,5187	3,67E-04	0,4128
SCGB1A1	Epithelial_cells	M - S	-2,9000	2,32E-03	0,9993
MT1X	Macrophage	HC - M	-1,3548	4,97E-03	0,9987
OSM	Macrophage	HC - S	1,0926	8,75E-04	0,9267
MT1G	Macrophage	M - S	2,4695	6,18E-05	0,1309
S100P	Monocyte	HC - M	-4,4580	1,01E-11	3,08E-08
WFDC2	Monocyte	HC - S	-3,1281	1,23E-09	8,45E-07
GSTA2	Monocyte	M - S	2,6629	2,78E-11	4,23E-08
CCL3	Pre-B_cell_CD34-	HC - M	1,1965	1,93E-04	0,0218
NUPR1	Pre-B_cell_CD34-	HC - S	-0,2746	4,13E-06	0,0038
GPNMB	Pre-B_cell_CD34-	M - S	-0,2023	4,12E-06	0,0038

Para visualização do efeito e significância estatística da análise DE realizada as Figuras 4.34 e 4.35 ilustram os gráficos *volcano plots* de acordo com a comparação e tipo de célula obtidos pelo modelos linear misto e edgeR, respectivamente. Como previamente discutido o modelo linear misto apresentou mais genes com maior significância estatística,

porém baixos logFC estimados, enquanto que o modelo edgeR apresentou resultados opostos, com altos logFC, porém poucos genes com significância estatística. Tais resultados revelam a complexidade de realizar a análise DE em scRNA-seq e evidenciam que o uso de diferentes modelos pode resultar conclusões distantes.

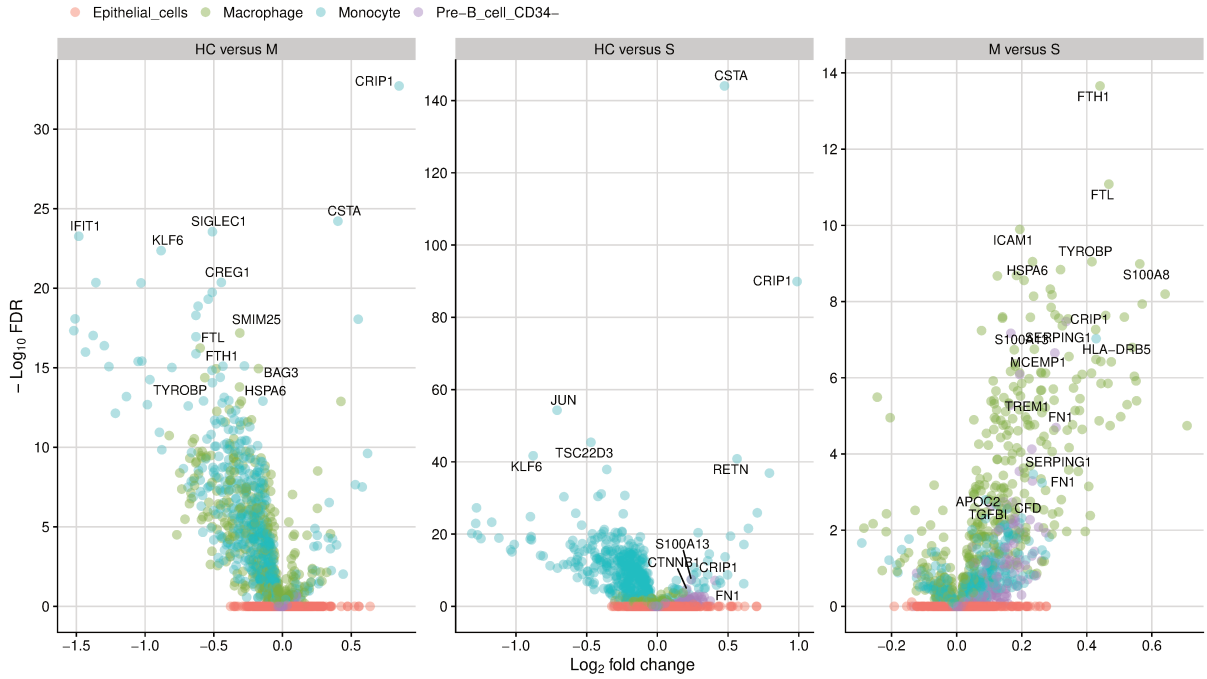


Figura 4.34 – *Volcano plots* comparando o log-fold change e FDR estimados para cada gene do modelo linear misto conforme o tipo de célula.

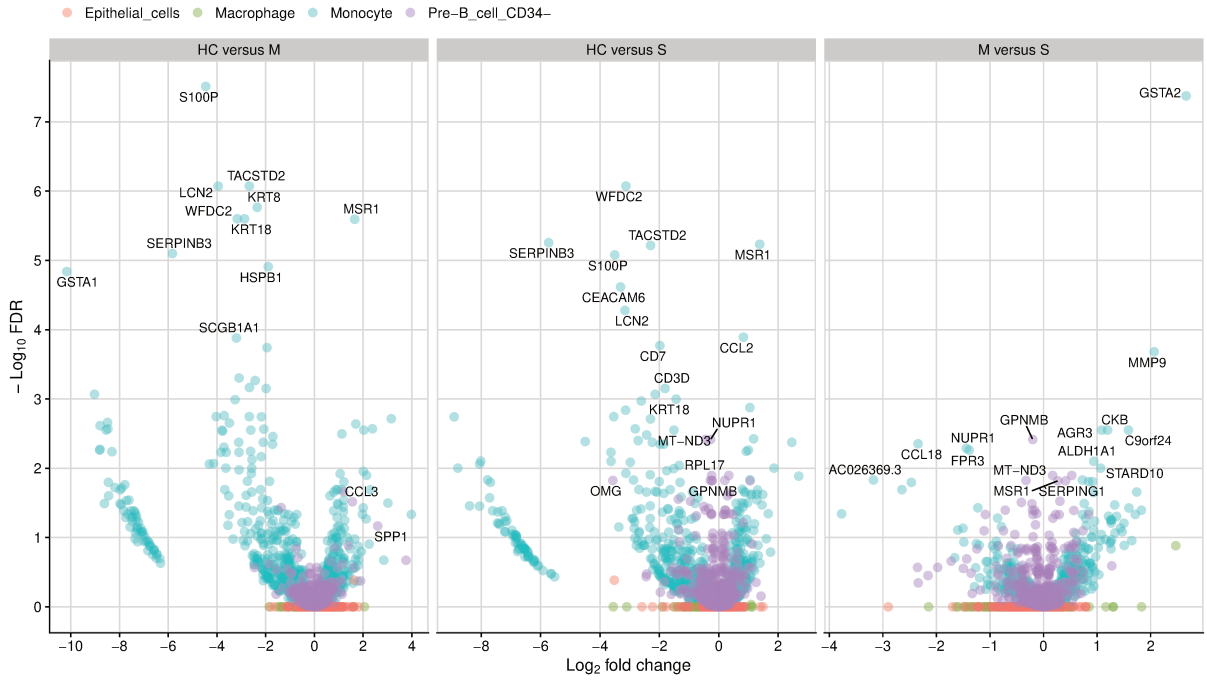


Figura 4.35 – *Volcano plots* comparando o log-fold change e FDR estimados para cada gene do modelo edgeR conforme o tipo de célula.

Nesse sentido, para validar as conclusões dos modelos é fundamental inspecionar graficamente as distribuições dos genes identificados como DE tanto no nível das células quanto no nível dos indivíduos. As Figuras 4.36 e 4.37 são exemplos de visualização das expressões gênicas no nível do indivíduo. Tais gráficos reportam a expressão gênica média dos top 6 genes com menor valor-p para cada indivíduo dentro de cada tipo de célula. Analisando a Figura 4.36, observa-se que nas células Epiteliais o gene CKS2 apresenta nível de expressão médio baixo nos grupos controle e moderado, em comparação com os indivíduos do grupo severo. Para as células do tipo monócito percebe-se que os genes CRP1, CSTA e HLA-DRB5 apresentam alta expressão média para os indivíduos do grupo controle em relação aos demais indivíduos. Nas células do tipo Macrófagos os genes FTH1, SMIM25, FTL e BAG3 apresentam expressão média alta nos indivíduos do grupo moderado em relação aos demais. Para as células Pre-B-CD34- o indivíduo HC2 apresentou somente uma célula, por isso o nível médio de expressão está diferente. Além disso, nota-se que para esta população de células todos os genes reportados apresentaram expressões médias altamente regulada para o grupo controle e com baixa regulação para os indivíduos com COVID-19 em estado severo.

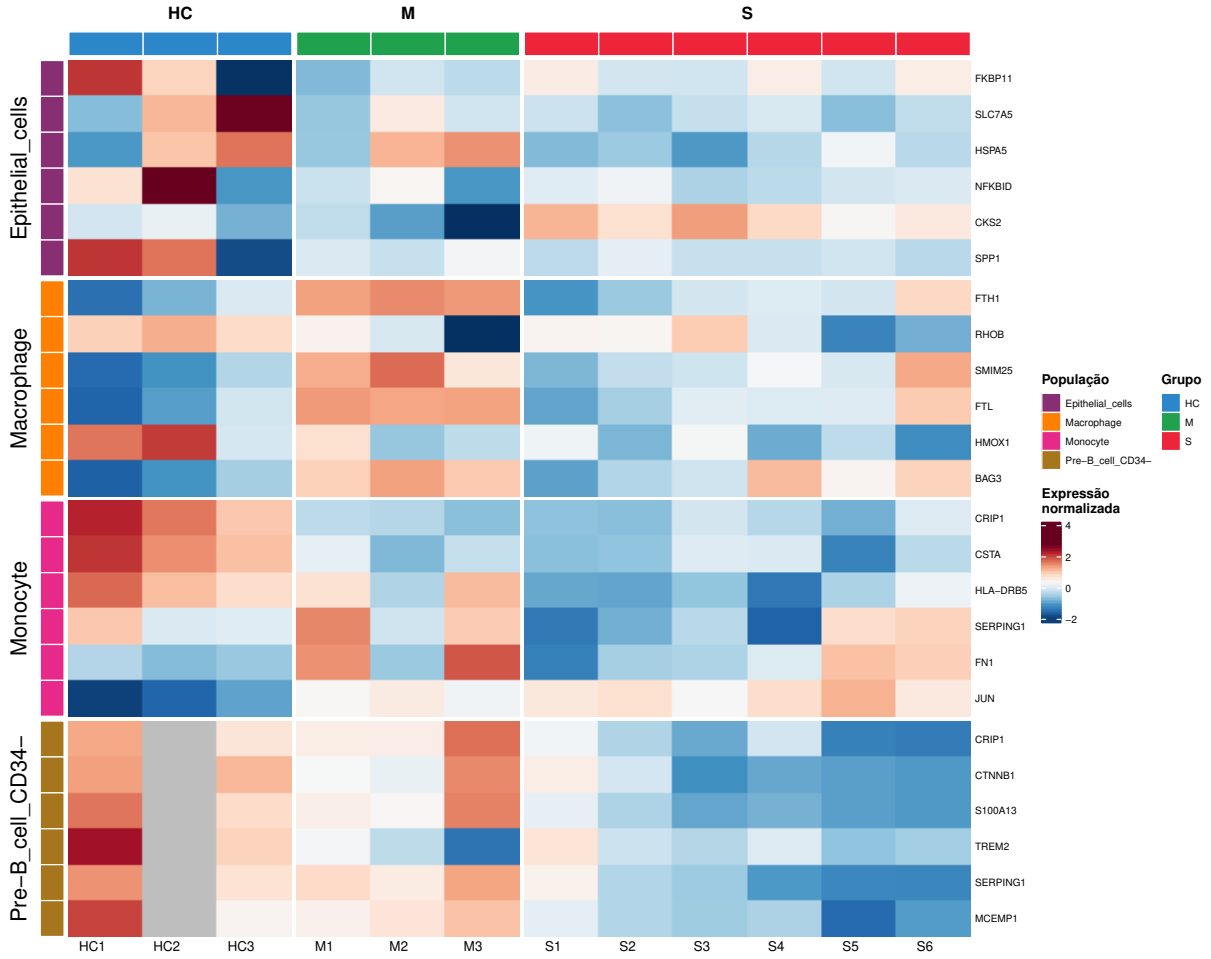


Figura 4.36 – Expressão média normalizada dos top 6 genes com menor valor-p identificados pelo modelo linear misto.

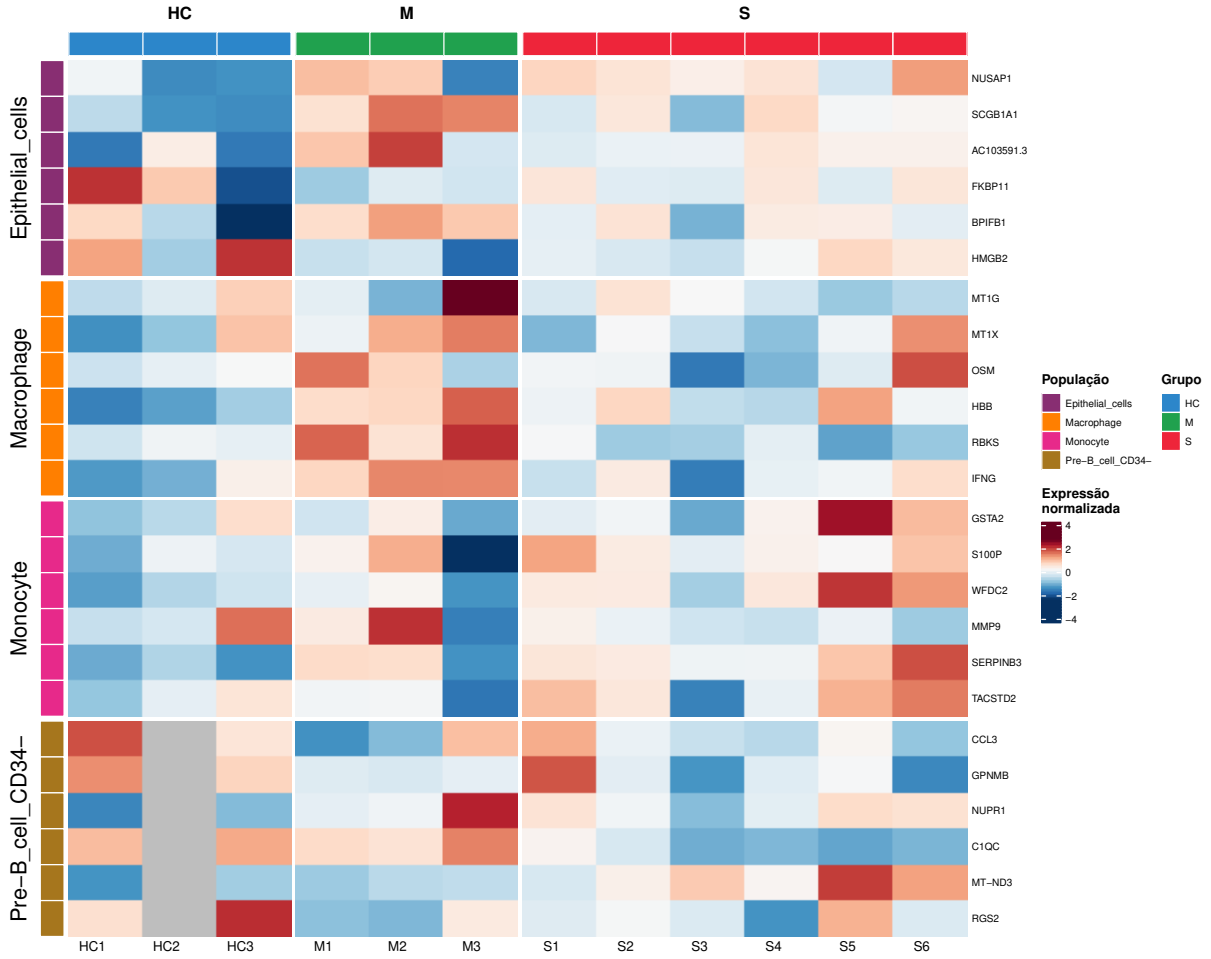


Figura 4.37 – Expressão média normalizada dos top 6 genes com menor valor-p identificados pelo modelo edgeR.

As Figuras 4.38, 4.39, 4.40, 4.41, 4.42, 4.43, 4.44 e 4.45 apresentam a distribuição dos top 6 genes identificados pelos modelos linear misto e edgeR. A possibilidade de investigar a distribuição das expressões gênicas no nível da células é uma das principais vantagens das tecnologias de scRNA-seq, sendo possível descobrir no nível mais granular (a célula) diferenças nas expressões gênicas entre os grupos de indivíduos e entre os próprios indivíduos. Por exemplo, analisando a Figura 4.40 verifica-se que existe uma concentração de células dos pacientes graves na cauda superior para os genes FN1 e HLA-DRB5 comparado com os grupos controle e moderado.

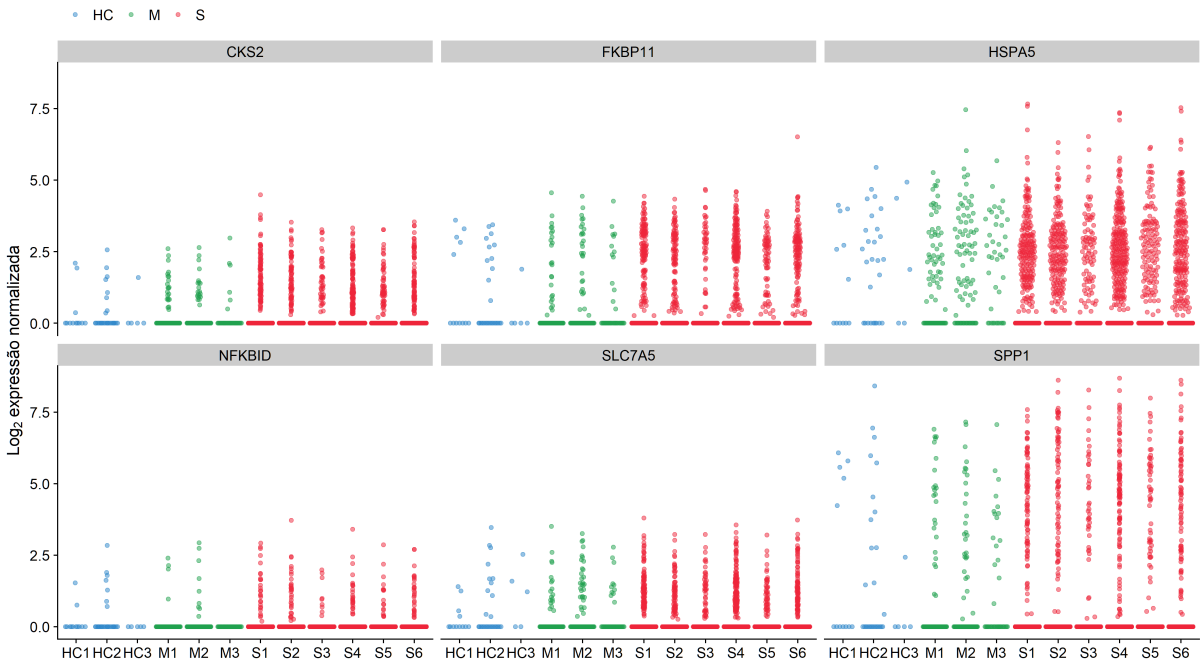


Figura 4.38 – Distribuição das expressões $\log_2$ normalizadas dos top 6 genes com menor valor-p identificados pelo modelo linear misto para as células epiteliais.

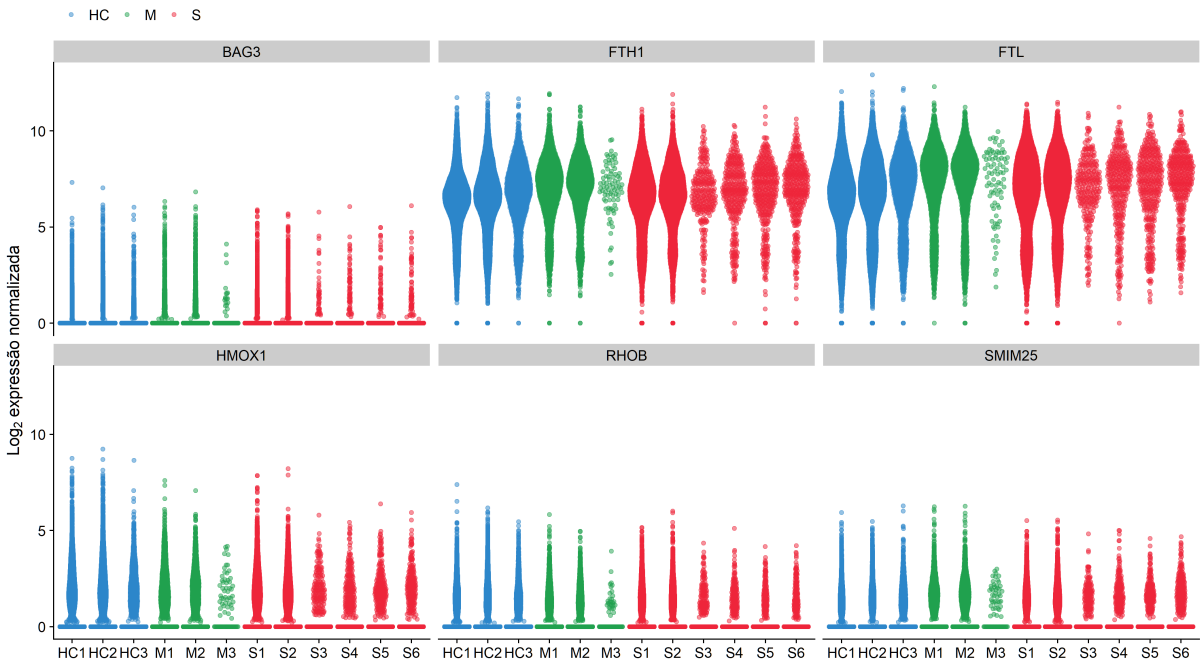


Figura 4.39 – Distribuição das expressões $\log_2$ normalizadas dos top 6 genes com menor valor-p identificados pelo modelo linear misto para as células macrófagos.

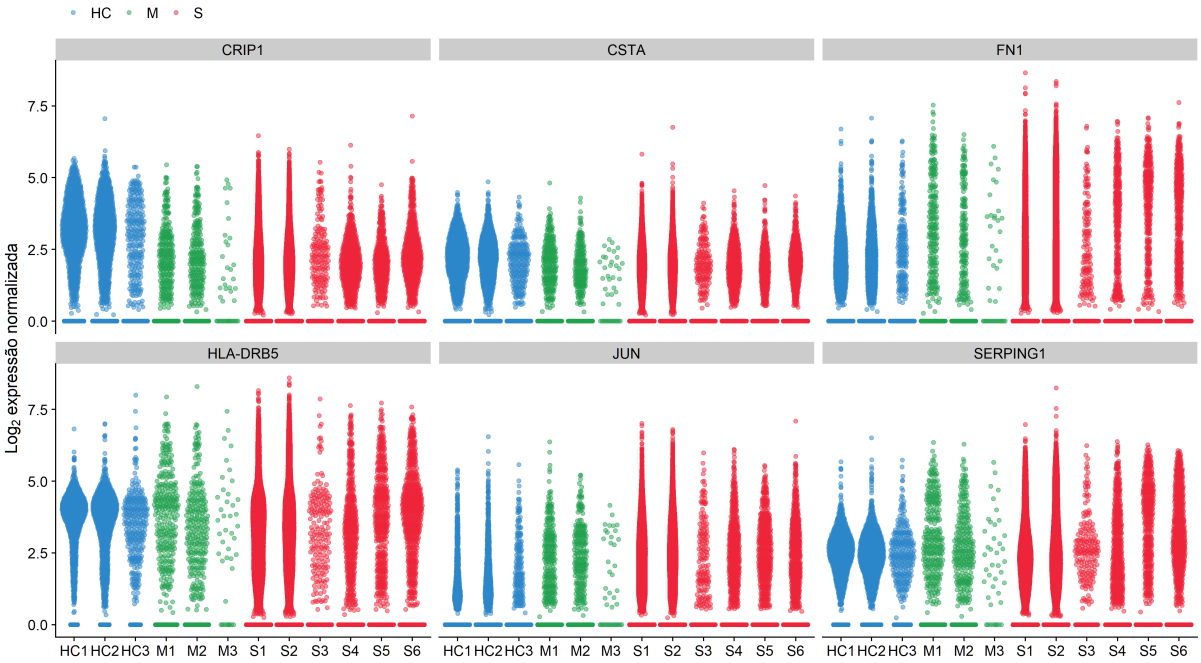


Figura 4.40 – Distribuição das expressões $\log_2$ normalizadas dos top 6 genes com menor valor-p identificados pelo modelo linear misto para as células monócitos.

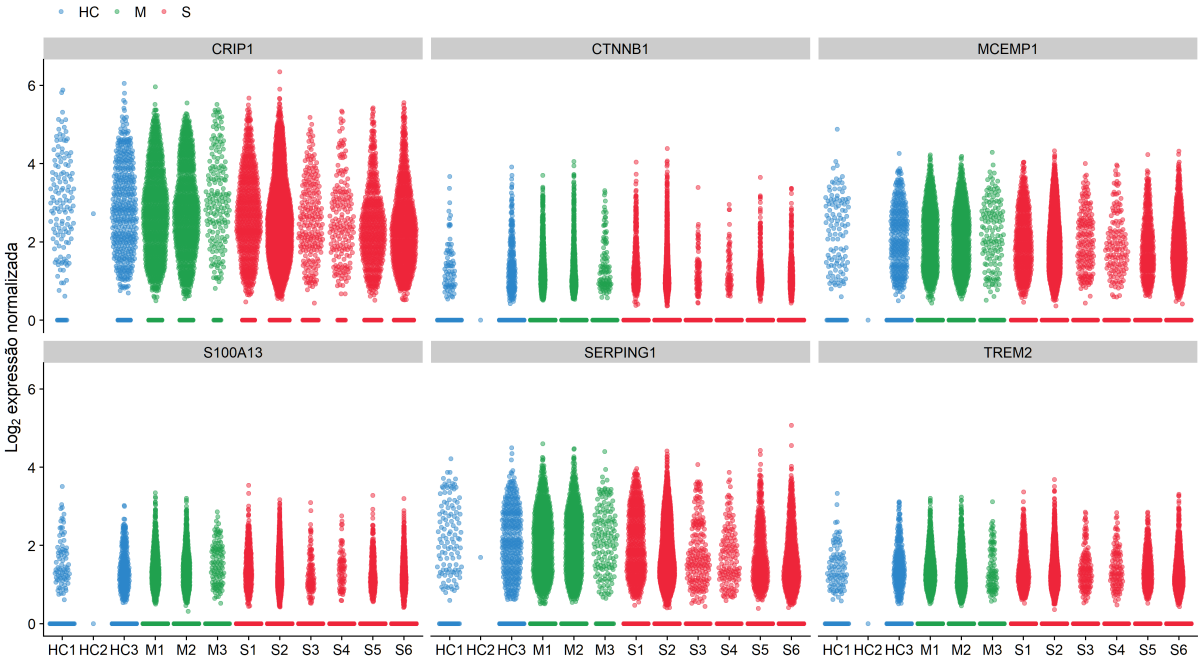


Figura 4.41 – Distribuição das expressões $\log_2$ normalizadas dos top 6 genes com menor valor-p identificados pelo modelo linear misto para as células Pre-B_cell_CD34-.

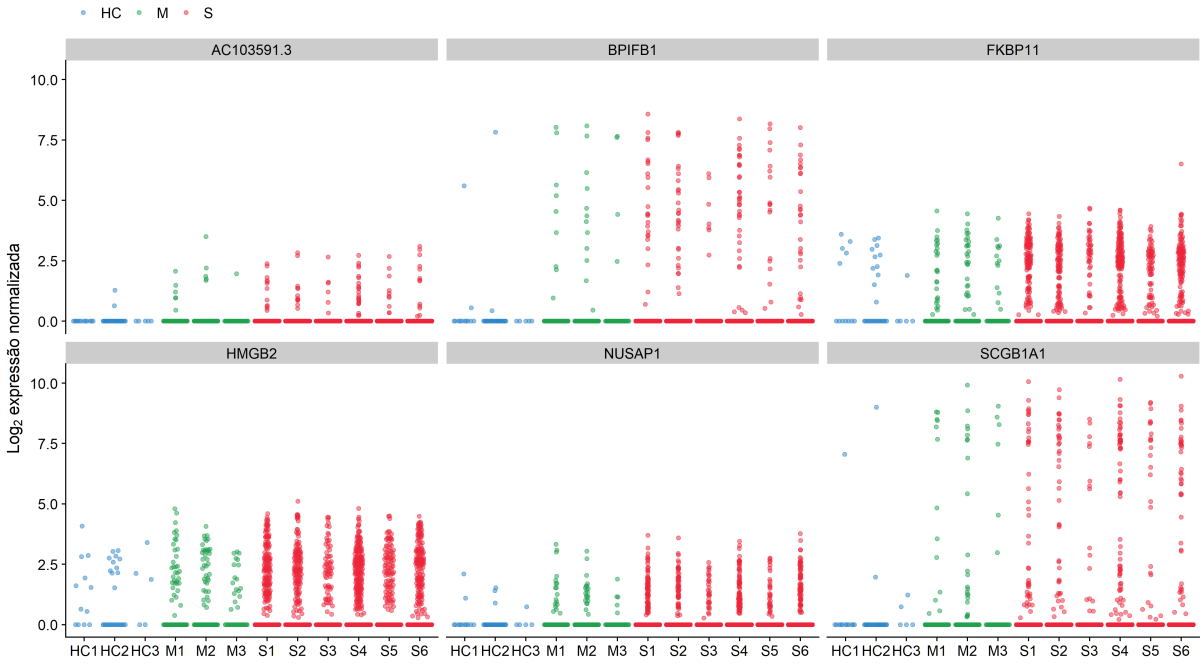


Figura 4.42 – Distribuição das expressões $\log_2$ normalizadas dos top 6 genes com menor valor-p identificados pelo modelo edgeR para as células epiteliais.

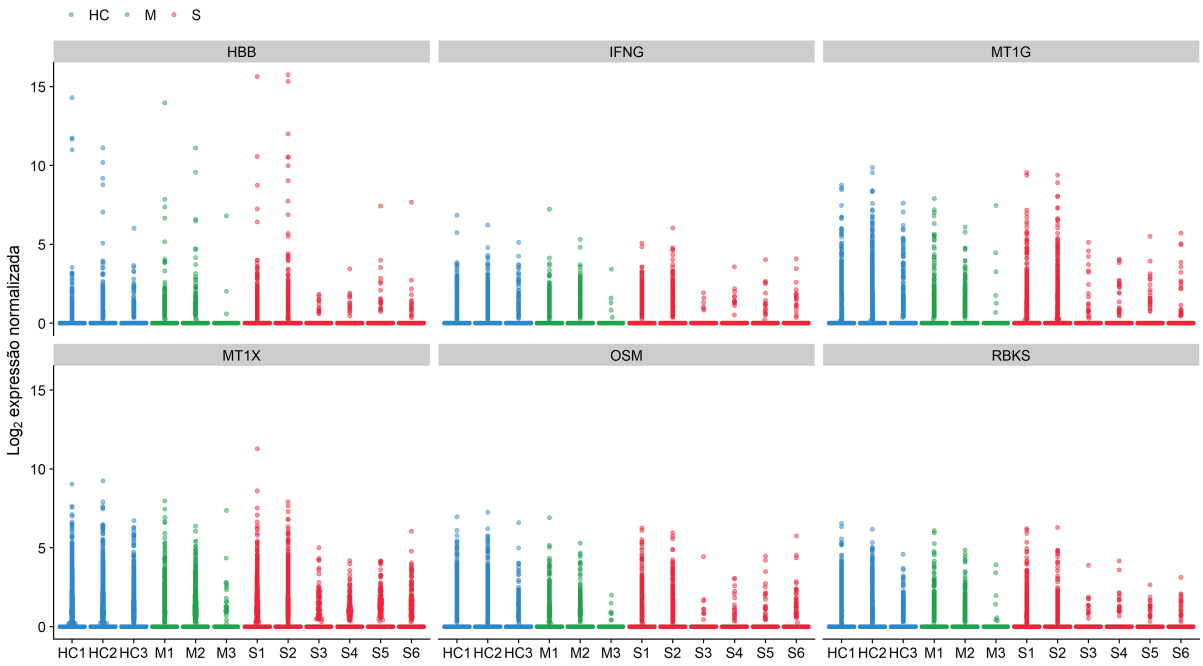


Figura 4.43 – Distribuição das expressões $\log_2$ normalizadas dos top 6 genes com menor valor-p identificados pelo modelo edgeR para as células macrófagos.

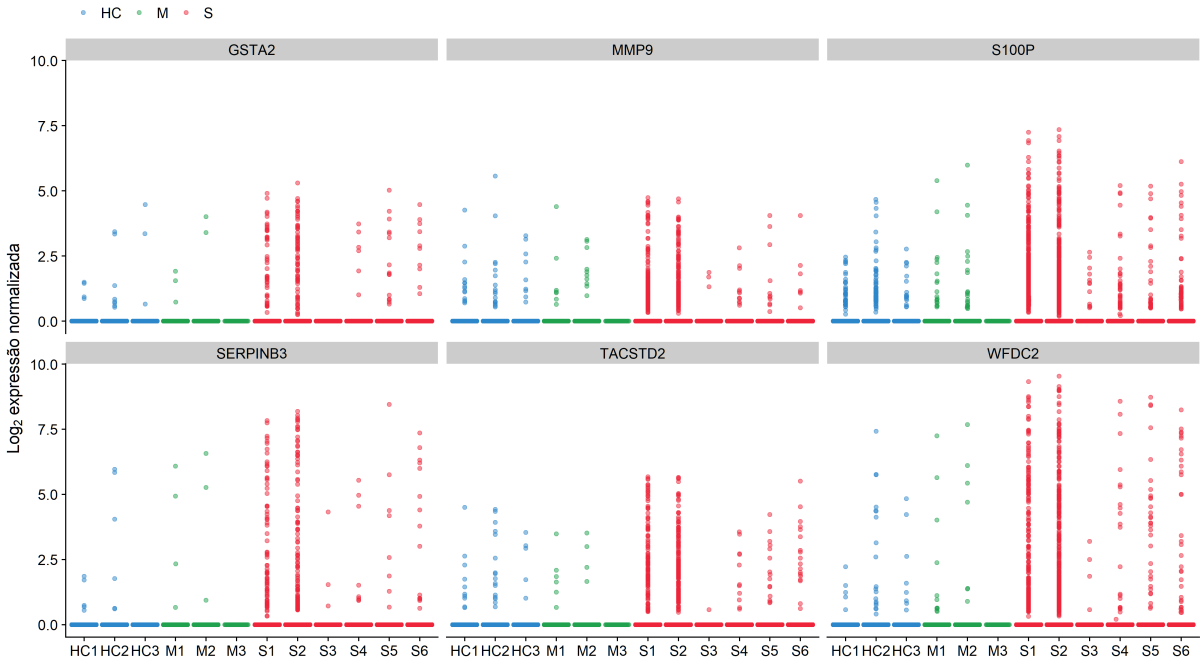


Figura 4.44 – Distribuição das expressões $\log_2$ normalizadas dos top 6 genes com menor valor-p identificados pelo modelo edgeR para as células monócitos.

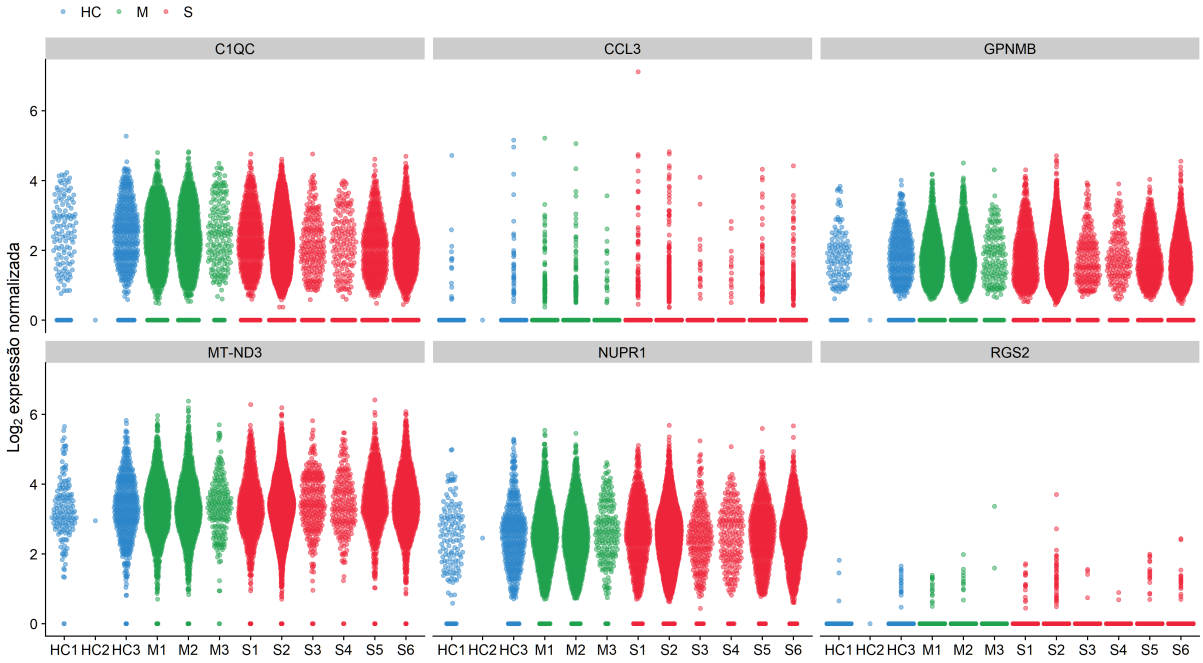


Figura 4.45 – Distribuição das expressões $\log_2$ normalizadas dos top 6 genes com menor valor-p identificados pelo edgeR para as células Pre-B_cell_CD34-.

5 Considerações finais

Esta pesquisa objetivou estudar as principais metodologias para análise de dados provenientes de experimentos scRNA-seq. Além disso, motivado pela estrutura experimental onde células de diferentes organismos são analisadas simultaneamente, realizou-se um estudo de simulação, incorporando características particulares dos dados de scRNA-seq, comparando os modelos edgeR e linear misto para análise da expressão gênica. Os resultados das simulações indicaram que o modelo linear misto com efeito aleatório da origem da célula apresentou performance satisfatória, controlando o erro do Tipo I, FDR e apresentando alto poder para detectar diferenças realmente verdadeiras entre grupos de células com expressões gênicas alteradas.

Por fim, uma análise dos dados reais das células BALF de doze indivíduos com e sem COVID-19 foi conduzida seguindo as metodologias discutidas ao longo da dissertação. Especificamente, análises exploratórias foram conduzidas para visualizar diferenças entre as células dos indivíduos, além disso, a análise de agrupamento foi realizada e treze grupos de células foram propostos, sendo anotados em cinco tipos de células distintos, que revelaram as particularidades biológicas dos grupos. Ainda na análise de dados reais os métodos utilizados no estudo de simulação foram aplicados para analisar a expressão gênica conforme a condição dos indivíduos. Diferenças entre os métodos foram observadas em relação aos genes detectados e a intensidade da diferença estimada (log-*fold change*). Tal diferença é justificada pela natureza da abordagem, enquanto que o modelo edgeR avalia diferença nas amostras pseudo *bulk* o modelo linear misto atua diretamente nas expressões gênicas no nível na célula e portanto possui maior sensibilidade. Análises gráficas também justificaram as diferenças observadas e permitiram acessar no nível mais granular possível, a célula, onde os genes são diferencialmente expressos, evidenciando o potencial dos experimentos scRNA-seq.

Como forma de reproduzir os resultados apresentados neste trabalho, as rotinas computacionais do estudo de simulação e análise de dados reais apresentados no Capítulo 4 estão disponíveis em um repositório público do GitHub (<https://github.com/AndrMenezes/dissertation>). O software estatístico R (R Core Team, 2020) foi utilizado extensivamente e vários pacotes do projeto Bioconductor (HUBER et al., 2015) foram considerados. Os códigos disponíveis no repositório contribuem para a pesquisa reproduzível e facilitam leitores interessados que desejam realizar novas análises e/ou comparações. Embora não apresente nenhuma inovação metodológica esta dissertação contribui, além da formação crítica como analista de dados do autor, como um material introdutório para a análise de dados em scRNA-seq na língua portuguesa, face a inexistência, em princípio, de outra referência na língua citada.

Referências

ALBERTS, B. et al. **The Molecular Biology of the Cell**. 6. ed. Garland Science, Taylor & Francis Group, 2015. Citado 3 vezes nas páginas 22, 23 e 100.

ALEXANDROV, L. B.; STRATTON, M. R. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. **Current Opinion in Genetics & Development**, v. 24, p. 52–60, 2014. ISSN 0959-437X. Cancer genomics. Citado na página 24.

ALQUICIRA-HERNANDEZ, J. et al. scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data. **Genome Biology**, v. 20, n. 264, 2019. Citado na página 65.

AMEZQUITA, R. A. et al. Orchestrating single-cell analysis with Bioconductor. **Nature Methods**, v. 17, p. 137–145, 2020. Citado 18 vezes nas páginas 10, 18, 19, 31, 32, 33, 36, 37, 40, 41, 42, 43, 47, 49, 57, 65, 93 e 116.

ARAN, D. et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. **Nature Immunology**, v. 20, p. 163–172, 2019. Citado 2 vezes nas páginas 18 e 65.

BAGNOLI, J. W. et al. Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq. **Nature Communications**, v. 9, n. 2937, 2018. Citado na página 30.

BATES, D. et al. Fitting linear mixed-effects models using lme4. **Journal of Statistical Software**, v. 67, n. 1, p. 1–48, 2015. Citado 2 vezes nas páginas 74 e 81.

BENJAMINI, Y.; HOCHBERG, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. **Journal of the Royal Statistical Society. Series B (Methodological)**, [Royal Statistical Society, Wiley], v. 57, n. 1, p. 289–300, 1995. Citado na página 123.

BERGER, R. L.; HSU, J. C. Bioequivalence trials, intersection-union tests and equivalence confidence sets. **Statistical Science**, v. 11, n. 4, p. 283–319, 11 1996. Citado na página 64.

BLAINEY, P. C. The future is now: single-cell genomics of bacteria and archaea. **FEMS Microbiology Reviews**, v. 37, n. 3, p. 407–427, 2013. Citado na página 24.

BRENNECKE, P. et al. Accounting for technical noise in single-cell RNA-seq experiments. **Nature Methods**, v. 10, p. 1093–1095, 2013. Citado na página 49.

BRYS, G.; HUBERT, M.; STRUYF, A. A robust measure of skewness. **Journal of Computational and Graphical Statistics**, v. 13, n. 4, p. 996–1017, 2004. Citado na página 39.

BUTLER, A. et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species. **Nature Biotechnology**, v. 36, n. 5, p. 411–420, 2018. Citado na página 65.

CHEN, Y.; LUN, A.; SMYTH, G. From reads to genes to pathways: differential expression analysis of RNA-seq experiments using rsubread and the edgeR quasi-likelihood pipeline. **F1000Research**, v. 5, n. 1438, 2016. Citado na página 122.

CHEN, Y.; LUN, A. T. L.; SMYTH, G. K. Differential expression analysis of complex RNA-seq experiments using edgeR. In _____. New York, 2014. cap. Statistical Analysis of Next Generation Sequence Data, p. 51–74. Citado 5 vezes nas páginas 69, 70, 71, 72 e 80.

CHEUNG, F. et al. Sequencing medicago truncatula expressed sequenced tags using 454 life sciences technology. **BMC Genomics**, v. 7, n. 272, 2006. Citado na página 17.

CLARK, D. P.; PAZDERNIK, N. J.; MCGEHEE, M. R. **Molecular Biology**. 3. ed. Elsevier, 2019. Citado 3 vezes nas páginas 10, 21 e 22.

CLEVELAND, W. S.; DEVLIN, S. J. Locally weighted regression: An approach to regression analysis by local fitting. **Journal of the American Statistical Association**, Taylor & Francis, v. 83, n. 403, p. 596–610, 1988. Citado na página 48.

COX, D. R.; REID, N. Parameter orthogonality and approximate conditional inference. **Journal of the Royal Statistical Society. Series B (Methodological)**, [Royal Statistical Society, Wiley], v. 49, n. 1, p. 1–39, 1987. Citado 2 vezes nas páginas 69 e 70.

CRICK, F. Central dogma of molecular biology. **Nature**, v. 227, p. 561–563, 1970. Citado 2 vezes nas páginas 17 e 22.

CROWELL, H. L. et al. muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. **Nature Communications**, v. 11, n. 6077, 2020. Citado 3 vezes nas páginas 19, 68 e 72.

DIGGLE, P. J. et al. **Analysis of Longitudinal Data**. Oxford University Press, 2002. Citado na página 73.

DING, J. et al. Systematic comparative analysis of single cell rna-sequencing methods. **bioRxiv**, Cold Spring Harbor Laboratory, 2019. Citado na página 29.

EFRON, B. et al. Empirical Bayes analysis of a microarray experiment. **Journal of the American Statistical Association**, Taylor & Francis, v. 96, n. 456, p. 1151–1160, 2001. Citado na página 70.

EMRICH, S. J. et al. Gene discovery and annotation using LCM-454 transcriptome sequencing. **Genome Research**, v. 17, n. 1, p. 69–73, 2007. Citado na página 17.

EVERITT, B. S.; DUNN, G. **Applied Multivariate Data Analysis**. 2nd. ed. United KingdomJohn Wiley & Sons, Ltd, 2001. Citado na página 52.

EVERITT, B. S.; HOTHORN, T. **An Introduction to Applied Multivariate Analysis with R**. United States of AmericaSpringer, 2011. Citado na página 52.

FINAK, G. et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. **Genome Biology**, v. 16, n. 278, 2015. Citado 3 vezes nas páginas 67, 79 e 93.

FISHER, R. A. **Statistical Methods for Research Workers**. Oliver and Boyd (Edinburgh), 1925. Citado na página 101.

FLUIDIGM. **Single-Cell Analysis**. 2020. <https://www.fluidigm.com/applications/single-cell-analysis>. Acessado em 24/04/2020. Citado na página 27.

FU, R. et al. clustifyr: An R package for automated single-cell RNA sequencing cluster classification. **F1000 Research**, 2019. Citado 3 vezes nas páginas 65, 66 e 107.

GIBBONS, J. D.; CHAKRABORTI, S. **Nonparametric Statistical Inference**. 5th. ed. Chapman & Hall/CRC, 2011. Citado na página 59.

GIERAHN, T. M. et al. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. **Nature Methods**, v. 14, p. 395–398, 2017. Citado na página 30.

GIESBRECHT, F. G.; BURNS, J. C. Two-stage analysis based on a mixed model: Large-sample asymptotic theory and small-sample simulation results. **Biometrics**, [Wiley, International Biometric Society], v. 41, n. 2, p. 477–486, 1985. Citado 3 vezes nas páginas 74, 75 e 121.

HAFEMEISTER, C.; SATIJA, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. **Genome Biology**, v. 20, n. 296, 2019. Citado 2 vezes nas páginas 47 e 72.

HARTIGAN, J. A.; WONG, M. A. Algorithm as 136: A k-means clustering algorithm. **Journal of the Royal Statistical Society. Series C (Applied Statistics)**, [Wiley, Royal Statistical Society], v. 28, n. 1, p. 100–108, 1979. Citado na página 54.

HASHIMSHONY, T. et al. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-seq. **Genome Biology**, v. 17, n. 77, 2016. Citado na página 30.

HASHIMSHONY, T. et al. CEL-Seq: single-cell RNA-seq by multiplexed linear amplification. **Cell Reports**, v. 2, n. 3, p. 666–673, 2012. Citado na página 30.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 2nd. ed. Springer, 2009. Citado 2 vezes nas páginas 52 e 53.

HICKS, S. C.; PENG, R. D. Elements and principles for characterizing variation between data analyses. **arXiv**, 2019. Citado na página 19.

HOCHGERNER, H. et al. STRT-seq-2i: dual-index 5' single cell and nucleus RNA-seq on an addressable microwell array. **Scientific Reports**, 2017. Citado na página 30.

HOFFMAN, G. E.; ROUSSOS, P. Dream: powerful differential expression analysis for repeated measures designs. **Bioinformatics**, 2020. Btaa687. Citado na página 72.

HUBER, P. J. **Robust statistics**. John Wiley, 1981. Citado na página 38.

HUBER, W. et al. Orchestrating high-throughput genomic analysis with Bioconductor. **Nature Methods**, 2015. Citado 3 vezes nas páginas 19, 32 e 135.

HUBERT, M.; Van der Veeken, S. Outlier detection for skewed data. **Journal of Chemometrics**, v. 22, p. 235–246, 2008. Citado 2 vezes nas páginas 38 e 39.

HUBERT, M.; VANDERVIEREN, E. An adjusted boxplot for skewed distributions. **Computational Statistics & Data Analysis**, v. 52, n. 12, p. 5186–5201, 2008. Citado na página 38.

ILUMINA. **An introduction to Next-Generation Sequencing Technology**. 2017. <https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf>. Acessado em 30/04/2020. Citado 3 vezes nas páginas 10, 30 e 31.

ISLAM, S. et al. Highly multiplexed and strand-specific single-cell RNA 5' end sequencing. **Nature Protocols**, v. 7, p. 813–828, 2012. Citado na página 30.

ISLAM, S. et al. Quantitative single-cell RNA-seq with unique molecular identifiers. **Nature Methods**, v. 11, p. 163–166, 2014. Citado 4 vezes nas páginas 28, 35, 37 e 94.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: A review. **ACM Computing Surveys**, v. 31, n. 3, p. 264–323, 1999. Citado na página 52.

JAITIN, D. A. et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. **Science**, v. 343, n. 6172, p. 776–779, 2014. Citado na página 30.

JAMES, G. et al. **An Introduction to Statistical Learning with Applications in R**. Springer, 2013. Citado na página 52.

JIANG, L. et al. Synthetic spike-in standards for RNA-seq experiments. **Genome Research**, v. 21, p. 1543–1551, 2011. Citado na página 36.

JOHNSON, R. A.; WICHERN, D. W. **Applied Multivariate Statistical Analysis**. 6th. ed. United States of AmericaPearson Prentice Hall, 2007. Citado na página 52.

KASSAMBARA, A. **Practical Guide to Cluster Analysis in R. Unsupervised Machine Learning**. 1st. ed. STHDA, 2017. (Multivariate Analysis 1). Citado na página 52.

KHARCHENKO, P. V.; SILBERSTEIN, L.; SCADDEN, D. T. Bayesian approach to single-cell differential expression analysis. **Nature Methods**, v. 11, n. 740–742, 2014. Citado na página 67.

KISELEV, V. Y.; ANDREWS, T. S.; HEMBERG, M. Challenges in unsupervised clustering of single- cell RNA- seq data. **Nature Reviews Genetics**, v. 20, p. 273–282, 2019. Citado 2 vezes nas páginas 18 e 53.

KISELEV, V. Y.; YIU, A.; HEMBERG, M. scmap: projection of single-cell RNA-seq data across data sets. **Nature Methods**, v. 15, n. 5, p. 359–362, 2018. Citado na página 65.

KLEIN, A. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. **Cell**, v. 161, n. 5, p. 1187–1201, 2015. Citado 2 vezes nas páginas 27 e 30.

KOLODZIEJCZYK, A. et al. The technology and biology of single-cell RNA sequencing. **Molecular Cell**, v. 58, n. 4, p. 610–620, 2015. Citado 3 vezes nas páginas 27, 28 e 36.

KORTHAUER, K. D. et al. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. **Genome Biology**, v. 17, n. 22, 2016. Citado na página 67.

KUKURBA, K. R.; MONTGOMERY, S. B. RNA sequencing and analysis. **Cold Spring Harb Protoc.**, v. 2015, p. 951–969, 2015. Citado na página 17.

KUMAR, R. M. et al. Deconstructing transcriptional heterogeneity in pluripotent stem cells. **Nature**, v. 516, p. 56–61, 2014. Citado na página 17.

KUZNETSOVA, A.; BROCKHOFF, P. B.; CHRISTENSEN, R. H. B. lmerTest package: Tests in linear mixed effects models. **Journal of Statistical Software, Articles**, v. 82, n. 13, p. 1–26, 2017. Citado na página 75.

LAFZI, A. et al. Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies. **Nature Protocols**, v. 13, p. 2742–2757, 2018. Citado 3 vezes nas páginas 27, 28 e 29.

LAHNEMANN, D. et al. Eleven grand challenges in single-cell data science. **Genome Biology**, v. 21, n. 3, 2020. Citado 2 vezes nas páginas 18 e 19.

LAW, C. W. et al. voom: precision weights unlocklinear model analysis tools for RNA-seq read counts. **Genome Biology**, v. 15, n. R29, 2014. Citado 4 vezes nas páginas 66, 72, 73 e 74.

LIAO, M. et al. Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. **Nature Medicine**, v. 26, p. 842–844, 2020. Citado 7 vezes nas páginas 20, 78, 80, 83, 93, 94 e 102.

LIU, R. et al. Why weight? Modelling sample and observational level variability improves power in RNA-seq analyses. **Nucleic Acids Research**, v. 43, n. 15, p. e97–e97, 2015. Citado na página 66.

LLOYD, S. Least squares quantization in PCM. **IEEE Transactions on Information Theory**, Institute of Electrical and Electronics Engineers (IEEE), v. 28, n. 2, p. 129–137, 1982. Citado na página 54.

LOVE, M. I.; HUBER, W.; ANDERS, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. **Genome Biology**, v. 15, n. 550, 2014. Citado 2 vezes nas páginas 66 e 75.

LOVE, M. I. et al. Tximeta: Reference sequence checksums for provenance identification in RNA-seq. **PLOS Computational Biology**, Public Library of Science, v. 16, n. 2, p. 1–13, 2020. Citado na página 33.

LUECKEN, M. D.; THEIS, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. **Molecular Systems Biology**, v. 15, n. 6, p. 1–23, 2019. Citado na página 33.

LUN, A. T. L. Overcoming systematic errors caused by log-transformation of normalized single-cell RNA sequencing data. **bioRxiv**, 2018. Citado 2 vezes nas páginas 46 e 47.

LUN, A. T. L. **bluster: Clustering Algorithms for Bioconductor**. 2020. R package version 1.0.0. Citado na página 57.

LUN, A. T. L.; BACH, K.; MARIONI, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. **Genome Biology**, v. 17, n. 75, 2016. Citado 8 vezes nas páginas 10, 18, 19, 43, 45, 80, 83 e 98.

LUN, A. T. L. et al. Assessing the reliability of spike-in normalization for analyses of single-cell RNA sequencing data. **Genome Research**, v. 27, p. 1795–1806, 2017. Citado na página 43.

LUN, A. T. L.; MARIONI, J. C. Overcoming confounding plate effects in differential expression analyses of single-cell RNA-seq data. **Biostatistics**, v. 18, n. 3, p. 451–464, 2017. Citado 9 vezes nas páginas 19, 68, 72, 78, 79, 84, 90, 92 e 122.

LUN, A. T. L.; MCCARTHY, D. J.; MARIONI, J. C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. **F1000Research**, v. 5, n. 2122, 2016. Citado 7 vezes nas páginas 20, 39, 57, 61, 64, 65 e 101.

LUN, A. T. L. et al. Emptydrops: Distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. **Genome Biology**, v. 20, n. 63, 2019. Citado na página 33.

LUN, A. T. L.; RISSO, D. **SingleCellExperiment: S4 Classes for Single Cell Data**. 2019. R package version 1.8.0. Citado na página 32.

LUO, Q.; ZHANG, H. Emergence of bias during the synthesis and amplification of cDNA for scRNA-seq. **Adv Exp Med Biol.**, v. 1068, p. 149–158, 2018. Citado 2 vezes nas páginas 35 e 36.

MABBOTT, N. et al. An expression atlas of human primary cells: inference of gene function from coexpression networks. **BMC Genomics**, v. 14, n. 632, 2013. Citado na página 107.

MACOSKO, E. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. **Cell**, v. 161, n. 5, p. 1202–1214, 2015. Citado 2 vezes nas páginas 27 e 30.

MAECHLER, M. et al. **robustbase: Basic Robust Statistics**. 2020. R package version 0.93-6. Disponível em: <http://robustbase.r-forge.r-project.org/>. Citado na página 38.

MANN, H. B.; WHITNEY, D. R. On a test of whether one of two random variables is stochastically larger than the other. **Annals of Mathematical Statistics**, The Institute of Mathematical Statistics, v. 18, n. 1, p. 50–60, 1947. Citado na página 58.

MASON, S. J.; GRAHAM, N. E. Areas beneath the relative operating characteristics (roc) and relative operating levels (rol) curves: Statistical significance and interpretation. **Quarterly Journal of the Royal Meteorological Society**, v. 128, n. 584, p. 2145–2166, 2002. Citado na página 59.

MCCARTHY, D. J. et al. Scater: pre-processing, quality control, normalisation and visualisation of single-cell RNA-seq data in R. **Bioinformatics**, v. 33, n. 8, p. 1179–1186, 2017. Citado 2 vezes nas páginas 38 e 39.

MCCARTHY, D. J.; CHEN, Y.; SMYTH, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. **Nucleic Acids Research**, v. 40, n. 10, p. 4288–4297, 2012. Citado 5 vezes nas páginas 20, 66, 68, 69 e 72.

MCCARTHY, D. J.; SMYTH, G. K. Testing significance relative to a fold-change threshold is a TREAT. **Bioinformatics**, v. 25, n. 6, p. 765–771, 2009. Citado 3 vezes nas páginas 60, 90 e 125.

MCINNES, L.; HEALY, J.; MELVILLE, J. UMAP: Uniform manifold approximation and projection for dimension reduction. **Preprint at *arXiv***, 2018. Citado na página 51.

MELSTED, P. et al. Modular and efficient pre-processing of single-cell RNA-seq. **bioRxiv**, Cold Spring Harbor Laboratory, 2019. Citado na página 33.

MERAD, M.; MARTIN, J. Pathological inflammation in patients with covid-19: a key role for monocytes and macrophages. **Nature Reviews Immunology**, v. 20, p. 355–362, 2020. Citado na página 112.

MEREU, E. et al. Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. **Nature Biotechnology**, 2020. Citado na página 29.

METHODS, N. Method of the year 2013. **Nature Methods**, v. 11, n. 1, 2013. Citado na página 17.

MILLIGAN, G. W.; COOPER, M. C. An examination of procedures for determining the number of clusters in a data set. **Psychometrika**, v. 50, n. 2, p. 159–179, 1985. Citado na página 55.

MORGAN, M. et al. **SummarizedExperiment: SummarizedExperiment container**. 2020. R package version 1.20.0. Disponível em: <https://bioconductor.org/packages/SummarizedExperiment>. Citado na página 32.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. **Journal of the Royal Statistical Society. Series A (General)**, [Royal Statistical Society, Wiley], v. 135, n. 3, p. 370–384, 1972. ISSN 00359238. Citado na página 69.

OLSEN, T. K.; BARYAWNO, N. Introduction to single-cell rna sequencing. **Current Protocols in Molecular Biology**, v. 122, n. e57, 2018. Citado na página 17.

PICELLI, S. et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. **Nature Methods**, v. 10, p. 1096–1098, 2013. Citado 2 vezes nas páginas 17 e 30.

PINHEIRO, J. C. **Topics in Mixed Effects Models**. Tese (Doutorado) — University of Wisconsin – Madison, 1994. Citado na página 74.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2020. Disponível em: <https://www.R-project.org/>. Citado na página 135.

RAMSKOLD, D. et al. Full-length mRNA-seq from single-cell levels of RNA and individual circulating tumor cells. **Nature Biotechnology**, v. 30, p. 777–782, 2012. Citado na página 30.

REGEV, A. et al. The human cell atlas. **bioRxiv**, Cold Spring Harbor Laboratory, 2017. Citado na página 18.

RENCHER, A. C. **Methods of Multivariate Analysis**. 2nd. ed. United States of AmericaJohn Wiley & Sons, Inc, 2002. Citado 2 vezes nas páginas 52 e 53.

REUTER, J. A.; SPACEK, D. V.; SNYDER, M. P. High-throughput sequencing technologies. **Molecular cell**, v. 58, n. 4, p. 586–597, 2015. Citado na página 17.

RISSO D.AND PERRAUDEAU, F. et al. A general and flexible method for signal extraction from single-cell RNA-seq data. **Nature Communications**, v. 9, n. 284, 2018. Citado na página 79.

RITCHIE, M. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. **Nucleic Acids Res.**, v. 43, n. e47, 2015. Citado 2 vezes nas páginas 61 e 66.

ROBINSON, M. D.; MCCARTHY, D. J.; SMYTH, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. **Bioinformatics**, v. 26, n. 1, p. 139–140, 2009. Citado 6 vezes nas páginas 20, 66, 80, 84, 120 e 122.

ROBINSON, M. D.; OSHLACK, A. A scaling normalization method for differential expression analysis of RNA-seq data. **Genome Biology**, v. 11, n. R25, 2010. Citado na página 122.

ROBINSON, M. D.; SMYTH, G. K. Moderated statistical tests for assessing differences in tag abundance. **Bioinformatics**, v. 23, n. 21, p. 2881–2887, 2007. Citado 3 vezes nas páginas 66, 69 e 71.

ROBINSON, M. D.; SMYTH, G. K. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. **Biostatistics**, v. 9, n. 2, p. 321–332, 2007. Citado 4 vezes nas páginas 66, 69, 70 e 72.

ROHATGI, V. K. **Statistical Inference**. John Willey & Sons, Inc, 1984. Citado na página 88.

SALIBA, A.-E. et al. Single-cell RNA-seq: advances and future challenges. **Nucleic Acids Research**, v. 42, n. 14, p. 8845–8860, 2014. Citado na página 27.

SASAGAWA, Y. et al. Quartz-Seq2: a high-throughput single-cell RNA sequencing method that effectively uses limited sequence reads. **Genome Biology**, v. 19, n. 28, 2018. Citado na página 30.

SASAGAWA, Y. et al. Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. **Genome Biology**, v. 14, n. 3097, 2013. Citado na página 30.

SCHENA, M. et al. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. **Science**, American Association for the Advancement of Science, v. 270, n. 5235, p. 467–470, 1995. Citado na página 66.

SCHURCH, N. J. et al. How many biological replicates are needed in an rna-seqexperiment and which differential expression tool shouldyou use? **RNA**, v. 22, n. 6, p. 839–851, 2016. Citado na página 67.

SEBER, G. A. F.; LEE, A. J. **Linear Regression Analysis**. 2th. ed. New York, 2003. Citado na página 62.

SERRATÌ, S. et al. Next-generation sequencing: advances and applications in cancer diagnosis. **Onco Targets Ther**, v. 9, p. 7355–7365, 2016. Citado na página 17.

SHALEK, A. K. et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. **Nature**, v. 498, p. 236–240, 2013. Citado na página 17.

SHAPIRO, E.; BIEZUNER, T.; LINNARSSON, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. **Nature Reviews Genetics**, v. 14, p. 618–630, 2013. Citado na página 27.

SHIRKHORSHIDI, A. S.; AGHABOZORGI, S.; WAH, T. Y. A comparison study on similarity and dissimilarity measures in clustering continuous data. **PLOS ONE**, Public Library of Science, v. 10, n. 12, p. 1–20, 2015. Citado na página 53.

SMITH, G. D.; EBRAHIM, S. Data dredging, bias, or confounding. **BMJ**, BMJ Publishing Group Ltd, v. 325, n. 7378, p. 1437–1438, 2002. Citado na página 114.

SMYTH, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. **Statistical Applications in Genetics and Molecular Biology**, v. 3, n. 1, 2004. Citado 3 vezes nas páginas 60, 66 e 70.

SMYTH, G. K. **limma: Linear Models for Microarray Data**. In: Gentleman R., Carey V.J., Huber W., Irizarry R.A., Dudoit S. (eds) Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Statistics for Biology and Health, 2005. Citado na página 66.

SMYTH, G. K. Optimization and nonlinear equations. **Encyclopedia of Biostatistics**, John Wiley and Sons, Ltd, p. 3088–3095, 2005. Citado na página 69.

SMYTH, G. K.; VERBYLA, A. P. Adjusted likelihood methods for modelling dispersion in generalized linear models. **Environmetrics**, v. 10, n. 6, p. 695–709, 1999. Citado na página 69.

SONESON, C.; ROBINSON, M. Bias, robustness and scalability in single-cell differential expression analysis. **Nature Methods**, v. 15, p. 255–261, 2018. Citado 4 vezes nas páginas 18, 19, 67 e 90.

SOUMILLON, M. et al. Characterization of directed differentiation by high-throughput single-cell RNA-seq. **bioRxiv**, 2014. Citado na página 30.

SRIVASTAVA, A. et al. Alevin efficiently estimates accurate gene abundances from dscRNA-seq data. **Genome Biol.**, v. 20, n. 65, 2019. Citado na página 33.

STOUFFER, S. et al. **The American Soldier, Vol. 1 - Adjustment during Army Life.** Princeton, 1949. Citado na página 61.

STUART, T. et al. Comprehensive integration of single-cell data. **Cell**, v. 177, p. 1888–1902, 2019. Citado na página 93.

SUVÀ, M. L.; TIROSH, I. Single-cell RNA sequencing in cancer: Lessons learned and emerging challenges. **Molecular Cell**, v. 75, n. 1, p. 7–12, 2019. Citado na página 24.

SVENSSON, V. Droplet scrna-seq is not zero-inflated. **Nature Biotechnology**, v. 38, p. 147–150, 2020. Citado 2 vezes nas páginas 47 e 79.

SVENSSON, V. et al. Power analysis of single-cell rna-sequencing experiments. **Nature Methods**, Nature Publishing Group, v. 14, p. 381–387, 2017. Citado na página 29.

TANG, F. et al. mRNA-Seq whole-transcriptome analysis of a single cell. **Nature Methods**, v. 6, p. 377–382, 2009. Citado 2 vezes nas páginas 17 e 24.

TIAN, L. et al. scPipe: A flexible R/Bioconductor preprocessing pipeline for single-cell RNA-sequencing data. **PLOS Computational Biology**, Public Library of Science, v. 14, n. 8, p. 1–15, 2018. Citado na página 33.

TIBSHIRANI, R.; WALTHER, G.; HASTIE, T. Estimating the number of clusters in a data set via the Gap statistic. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, v. 63, n. 2, p. 411–423, 2002. Citado 2 vezes nas páginas 55 e 107.

TIROSH, I.; SUVÀ, M. L. Deciphering human tumor biology by single-cell expression profiling. **Annual Review of Cancer Biology**, v. 3, n. 1, p. 151–166, 2019. Citado na página 24.

TOWNES, F. W. et al. Feature selection and dimension reduction for single-cell RNA-seq based on a multinomial model. **Genome Biology**, v. 20, n. 295, 2019. Citado 3 vezes nas páginas 47, 48 e 79.

TRAN, H. T. N. et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. **Genome Biology**, n. 12, 2020. Citado na página 102.

TRAPNELL, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. **Nature Biotechnology**, v. 32, p. 381–386, 2014. Citado na página 67.

VALLEJOS, C.; RICHARDSON, S.; MARIONI, J. Beyond comparisons of means: understanding changes in gene expression at the single-cell level. **Genome Biology**, v. 70, n. 17, 2016. Citado na página 67.

VALLEJOS, C. et al. Normalizing single-cell RNA sequencing data: challenges and opportunities. **Nature Methods**, v. 14, p. 565–571, 2017. Citado na página 18.

VALLEJOS, C. A.; MARIONI, J. C.; RICHARDSON, S. BASiCS: Bayesian analysis of single-cell sequencing data. **Computational Biology**, v. 11, n. e1004333, 2015. Citado 3 vezes nas páginas 36, 67 e 79.

Van der Maaten, L.; HINTON, G. Visualizing data using t-SNE. **Journal of Machine Learning Research**, v. 9, p. 2579–2605, 2008. Citado na página 51.

WANG, T. et al. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. **BMC Bioinformatics**, v. 20, n. 40, 2019. Citado na página 67.

WANG, X. Approximating bayesian inference by weighted likelihood. **The Canadian Journal of Statistics**, [Statistical Society of Canada, Wiley], v. 34, n. 2, p. 279–298, 2006. Citado na página 70.

WANG, Z.; GERSTEIN, M.; SNYDER, M. RNA-Seq: a revolutionary tool for transcriptomics. **Nature Reviews Genetics**, v. 10, p. 57–63, 2009. Citado 2 vezes nas páginas 17 e 24.

WANG, Z. et al. scruff: An R/Bioconductor package for preprocessing single-cell RNA-sequencing data. **BMC Bioinformatics**, v. 20, n. 222, 2019. Citado na página 33.

WEBER, A. P. et al. Sampling the arabidopsis transcriptome with massively parallel pyrosequencing. **Plant Physiology**, American Society of Plant Biologists, v. 144, n. 1, p. 32–42, 2007. Citado na página 17.

WHITLOCK, M. C. Combining probability from independent tests: The weighted Z-method is superior to Fisher's approach. **Journal of Evolutionary Biology**, v. 18, n. 5, p. 1368–1373, 2005. Citado na página 61.

YIP, S. H.; SHAM, P. C.; WANG, J. Evaluation of tools for highly variable gene discovery from single-cell RNA-seq data. **Briefings in Bioinformatics**, v. 20, p. 1583–1589, 2018. Citado na página 48.

ZAPPIA, L.; PHIPSON, B.; OSHLACK, A. Splatter: simulation of single-cell RNA sequencing data. **Genome Biology**, v. 18, n. 174, 2017. Citado na página 79.

ZAPPIA, L.; PHIPSON, B.; OSHLACK, A. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. **PLOS Computational Biology**, Public Library of Science, v. 14, n. 6, p. 1–14, 2018. Citado na página 19.

ZEISEL, A. et al. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. **Science**, v. 347, n. 6226, p. 1138–1142, 2015. Citado 4 vezes nas páginas 10, 40, 41 e 42.

ZHANG, X. et al. Comparative analysis of droplet-based ultra-high-throughput single-cell RNA-seq systems. **Molecular Cell**, v. 73, p. 130–142, 2019. Citado 3 vezes nas páginas 10, 27 e 28.

ZHENG, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. **Nature Communications**, v. 8, n. 14049, 2017. Citado 4 vezes nas páginas 17, 27, 30 e 33.

ZIEGENHAIN, C. et al. Comparative analysis of single-cell RNA sequencing methods. **Molecular Cell**, v. 65, n. 4, p. 631–643.e4, 2017. Citado na página 29.